**A Study on the Major Determining Factors for Box Office Success in the Film Industry**

**CSCI 4502/5502: Fall 2021**

| **Connor Dixon**  **University of Colorado Boulder**  **Boulder. Colorado, United States**  **codi2763@colorado.edu** | **Kaylee Engelhardt University of Colorado Boulder**  **Boulder. Colorado, United States**  **anen3020@colorado.edu** | **Baci Brunet**  **University of Colorado Boulder**  **Boulder. Colorado, United States**  **sebr8260@colorado.edu** | **Aidan Aarts**  **University of Colorado Boulder**  **Boulder. Colorado, United States aiaa4302@colorado.edu** |
| --- | --- | --- | --- |

**1 Introduction**

Predicting box office income is a critical issue in the film industry since it influences producers' and investors' financial decisions. In most cases, fundamental statistical approaches are used to make these predictions. While these methods are widely used, they frequently only provide a rough estimate of revenue before a film is released. The purpose of this research is to create a computational model for predicting box office revenues using public data from popular internet movie databases.

To achieve a successful model, this team will analyze and normalize past box office data. Using that data, this team will evaluate trends and patterns to predict performance of upcoming films. This prediction will be used to determine whether or not the films will be prudent and wise investments.

Movies can pose massive initial investments that need to be evaluated using a multitude of variables to determine the success and payout come opening day. Movies like Avengers Endgame have budgets of 356 million dollars, and box office profits of 2.45 billion. The model will consider the variables that affect these payouts in order to predict the profit a company can expect to make off of the film and initial investment.

Many of these firms already have systems in place to predict this type of behavior; but, if this study used datasets/external elements that the aforementioned systems don't, it could generate fresh ideas that the companies could implement later. At the very least, this research may reveal data sets that organizations haven't utilized or given enough attention to, allowing them to better target/improve their tactics.

**2 Literature Review**

Related work in box office performance expectations includes a paper by Eon Smit and N. A. Pangarker[1] on the various important factors on box office success for various movies. This paper performs a literature review on many articles printed before 2013. These articles often include primarily movies made for and in the United States of America.

The factors included in the first paper studied are “the creative sphere, the scheduling and release pattern, and the marketing effort”. In the creative sphere, this includes factors such as: whether the story is genuine, “cast, director, production budget, and rating.” For the marketing and release sphere, access to a major distributor is most important, followed by releasing films at peak viewing times during the year. For the marketing effort, the factors that are the most important are the marketing campaign size before release and word of mouth after release. Contrary to the opinion at the time, the author of this paper believed in the value of critics’ reviews of movies.

In the second paper studied by Eon Smit and N. A. Pangarker, the authors believed there to be too many extreme factors to account for, for success. Instead, they modelled the success of previous movies on several common factors and created probability distributions to determine the range and expected return for that type of movie. The factors that they divided movies based on are: sequels, genres, ratings, stars, budgets, and opening screens.” They believe individual stars are very important for a film’s success.

The third paper, by Eon Smit and N. A. Pangarker, attempted to account for the value of a movie based on the early box office results. This approach was chosen to help movie exhibitions decide what to show and when based on the early appeal. They focused primarily on the attendance numbers with little data on revenue.

Most other papers researched by Eon Smit and N. A. Pangarker attempt to estimate performance based on a variety of factors much like in the first paper that was researched. Overall, these papers attempted to study the value of stars and sequels primarily, but occasionally other topics.

In a paper by Jia Xiao, Xin Li, et al.[2], the authors explore the connection between movies’ trailers and the number of comments posted about them with the success of the movie at the actual release to try and determine the likely success of the movie. This paper analyzes the distribution of revenue from 2012 to 2015, so the dataset is rather small temporally, so not very accurate for this project for this class because temporal data was looked at. The other factors that the authors attempted to account for are investment, title length, script length, movie length, release schedule, rival, genre, cast, awards, and advertising.

A paper on success of international movies in Russia[3] covered factors such as: “distribution related (e.g., budget and franchise), brand and star effects (e.g., top actors or directors), and evaluation sources (e.g., critics and audience rating)” as well as a variety of cross-cultural factors such as “seasonality, time span between the world and local release, attendance of international stars at Russian movie premieres, and title adaptation to Russian culture.” They found that without accounting for the cross-cultural factors, movie success was nearly unpredictable. This shows that the project for this class will likely only be applicable for movies created for and by the United States.

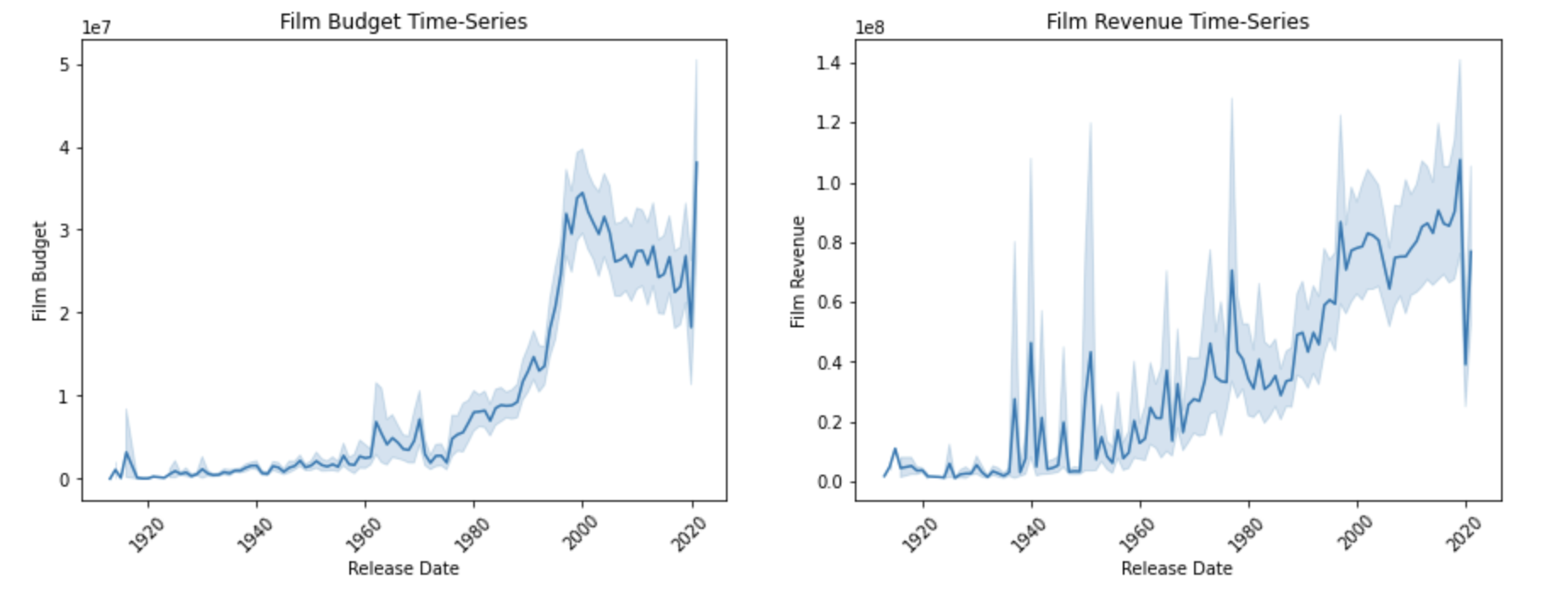
**3 Work Done**

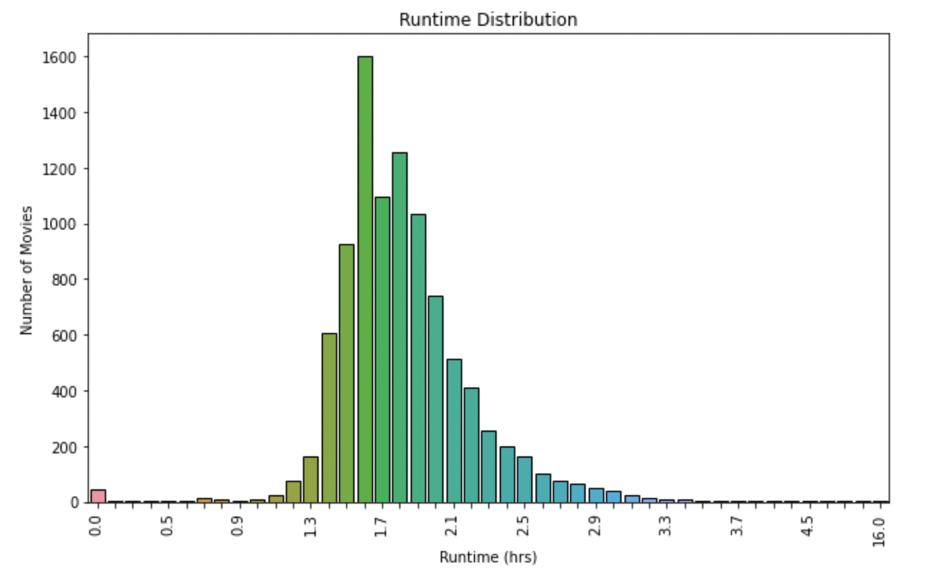
**3.1 Data Exploration**

​​The Movie Database (TMDB) is the source of box office data for this project. The information obtained using an API allowed the creation of the dataset used which consists of the top 9,971 movies, ranked by revenue in descending order. There are 12 attributes for each of those movies, which include the film title, gross revenue, budget, release date, associated genres, film popularity, original language, production companies involved, id (a unique identifier used to call the API in case the team want to gather more attributes), runtime, age rating, cast of film, and finally the production crew members.

First and foremost, one of the most important attributes in the dataset is budget, the set contains movies with a budget range of ($0, $380,000,000) the mean budget for movies in this range is approximately $2,818,970.

The attribute in which training to predict revenue is occurring has a range of ($671,565, $2,847,246,203) with a mean of approximately $67,042,032.

Shown above the distribution of movies release date can be seen corresponding to their budget and final revenue. The time range attribute only considers movies from within 1913 to 2020 and could potentially be broken up into smaller “eras” of movies to make the model more accurate given a certain time period.

​​

Shown above is the distribution of the runtime attribute which has a range from 11 minutes up until 960 minutes. Movies within this range have a mean runtime of 110 minutes and a mode of 100 minutes.

The original dataset contained an attribute for the language in which a movie was released in. After inspection the distribution of languages across the set was around 75% in English, 2% in Spanish, French, Hindi, and Japanese, and 1% in Korean, Dutch, Italian, Russian, and Chinese. Because English makes up such a large majority of movies in the dataset it might not be beneficial to even include this in the predictive model. That being said it may also be advantageous to only consider movies who’s original language is listed as English unless that would jeopardize training.

**3.1.1 Data Quality**

First and foremost, the data was checked for completeness. As mentioned earlier, the original data set consisted of 9971 records pre-cleaning. The following table describes the percentage of incomplete/null entries per attribute:

| title | 0.000000 |
| --- | --- |
| revenue | 0.000000 |
| budget | 0.000000 |
| release\_date | 0.401163 |
| genres | 0.752181 |
| popularity | 0.000000 |
| original\_language | 0.000000 |
| production\_companies | 3.841139 |
| id | 0.000000 |
| runtime | 0.160465 |
| cast | 0.531541 |
| crew | 0.671949 |

Per the table, the highest number of incomplete entries appears in the ‘production\_companies’ field, followed by ‘genres’, ‘cast’, ‘crew’, and ‘runtime’. The number of records containing null values totaled to 417 or 4% of the original set. Since none of the null entries for the above attributes could be inferred from any other combination of attributes in the data set, it was decided to drop these records from the dataset. After cleaning there were 9554 records left.

**3.1.2 Data Preprocessing and Preparation**

When this data was pulled from The Movie Database’s API, the attributes ‘cast’ and ‘crew’ were formatted as JSON objects. The aforementioned attributes contained each actor and crew member for a given film along with their associated popularity score (provided by The Movie Database). It was decided to split these attributes into a list of actor/crew members for each film and their associated average popularity scores (across each list), resulting in two new attributes, namely, ‘cast\_average’ and ‘crew\_average’ -- the mean value for those attributes are 2.35 and ~1 respectively).

The following attributes were formatted as lists and thus required encoding for further analysis: ‘production\_companies’, ‘genres’, ‘cast’, ‘crew' and ‘original\_language’. For ‘genre’ and ‘original\_language’ there are only 19 and 50 unique values, respectively, so the team went with a one-hot encoding. For ‘production\_companies’, ‘cast’ and ‘crew’ there were over 10,000 unique values. Up until now, it was decided that the best way to navigate this issue is by one-hot encoding the top 50 most frequently appearing companies, and 500 most frequently occurring actors and crew members.

Finally, the release date attribute was binned into 104 bins -- accounting for the years between 1918 and 2020 -- in order to encode it for use in our model training and testing.

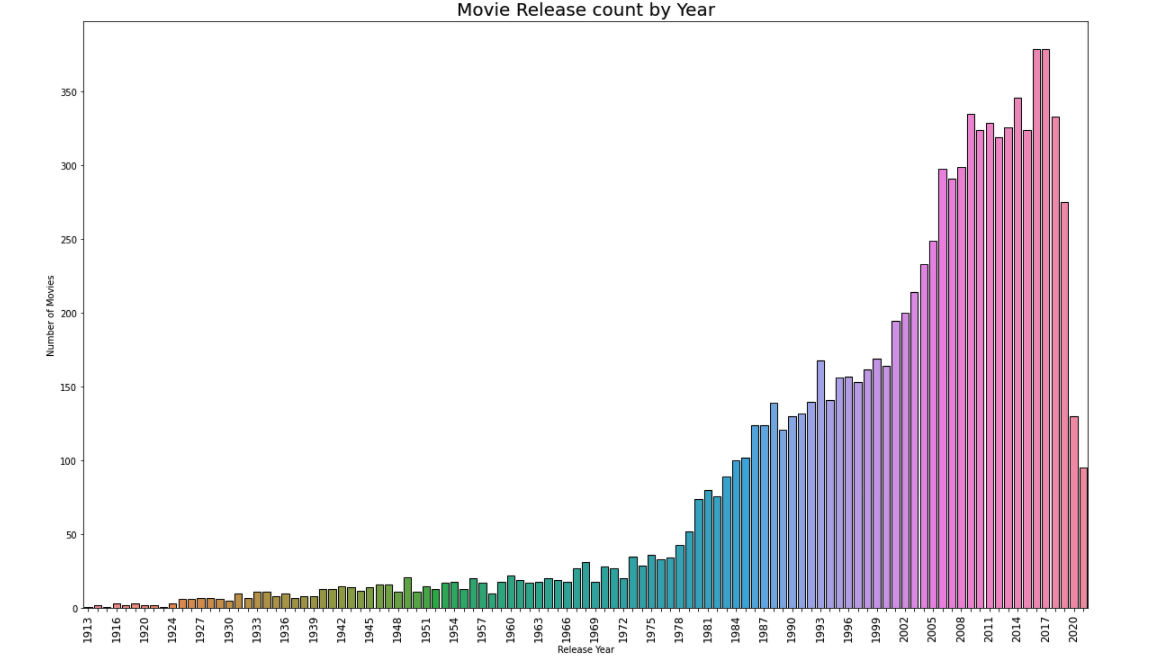
**3.1.3 Data Collection and Storage**

Once this data was collected, preprocessed and prepared, in order for the models to work as expected, and further processing to take place, the data needed to be stored. Initially, the data (both raw and preprocessed) was stored in a **Linode-hosted Cloud MySQL Database**, in order to allow for the easiest organization, storage and querying of the aforementioned data.

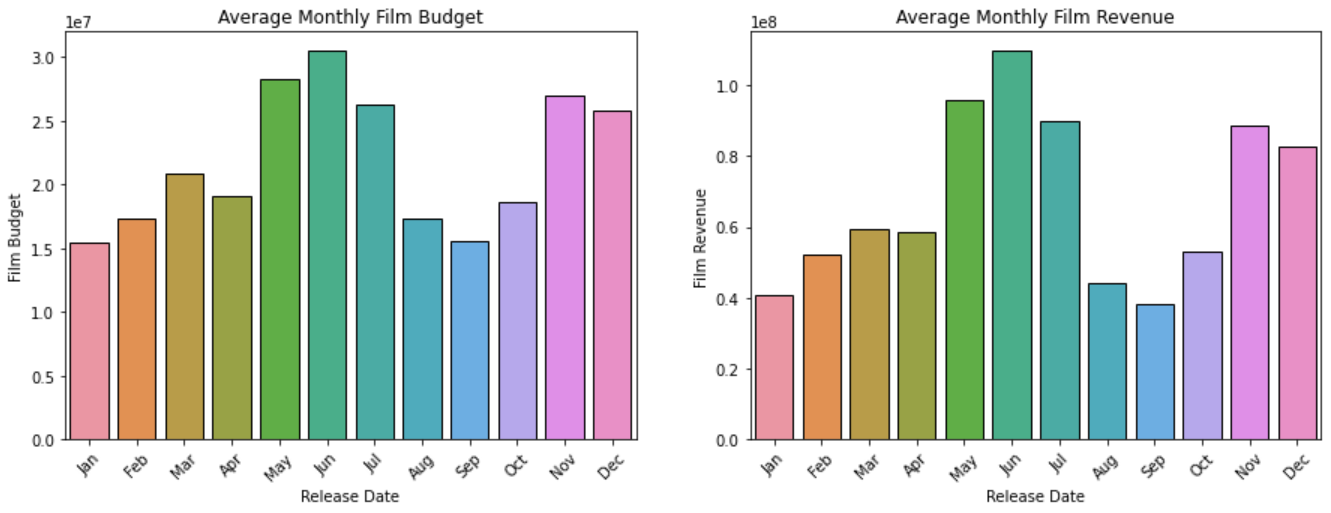
However, it was soon discovered that constantly querying a cloud hosted database via API calls was quite inefficient, so much so that it soon nullified the purpose for storing and organizing it in the database in the first place. Therefore, the data was then transferred to a locally stored collection of **.csv** files, which allowed for much quicker and efficient analysis and access by both the exploratory plot libraries and the prediction models.

Thus, with this system in place, and since the models and plots were able to access the data locally rather than over a server connection with tedious API calls, the performance of these models improved exponentially, which is crucial when dealing with both datasets of this size and models of this complexity.

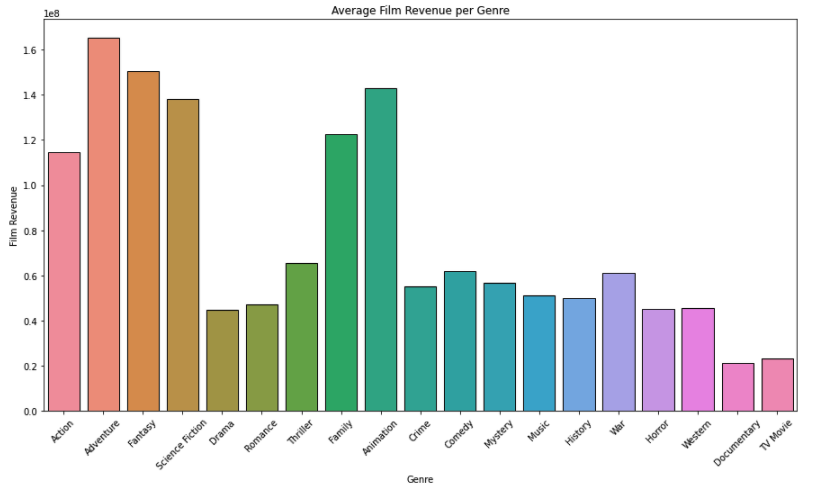
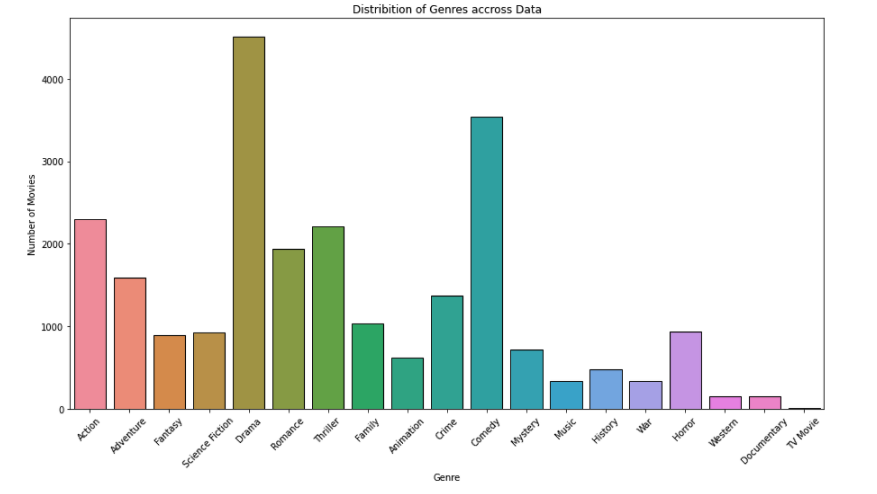
**3.1.4 Exploratory Analysis**



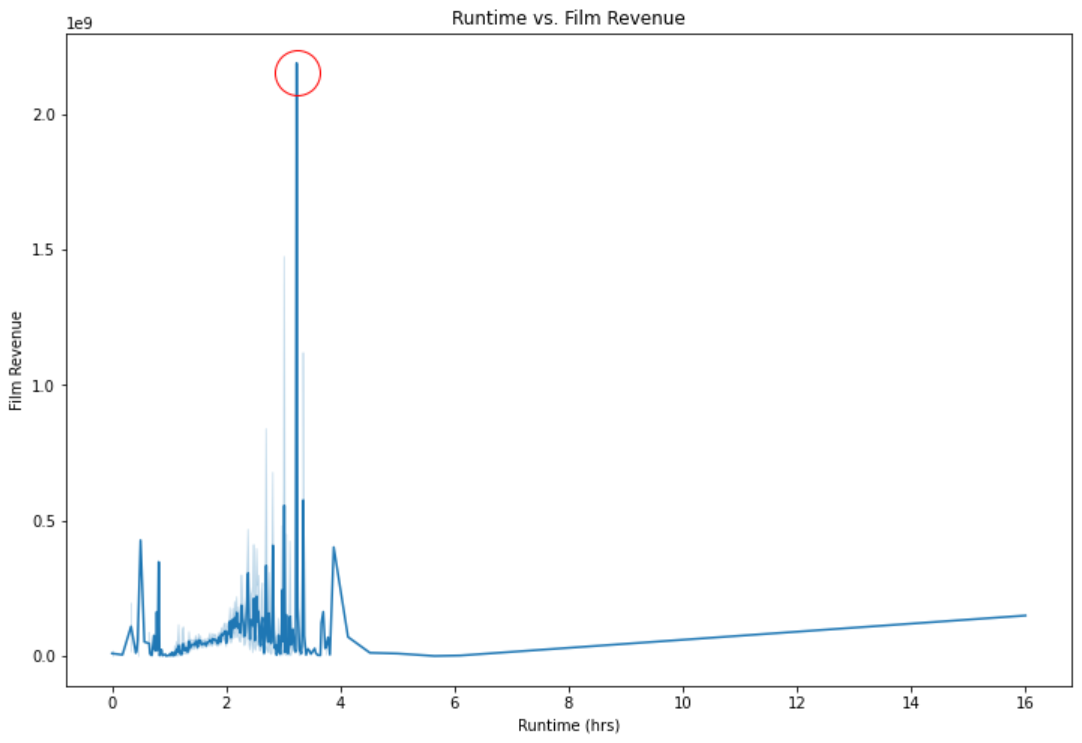
The above plot shows the distribution of movies with regards to the year in which they were released. It can be seen that the majority of the films in the data set were released post 1980. Furthermore, around 50% of the films in the data set were released after the year 2000. This is important to note because this project will be trying to predict the revenue for upcoming films, so it must be ensured that the data is not skewed toward older films.



These graphs show how film budget and revenue change depending on which month of the year the movie is released. From here it can be seen that the most budget heavy and highest grossing movies are generally released in summer and winter months. It is important to note that because of the heavy correlation between budget and revenue it is possible that the highest budget films are released in the summer and winter and therefore the highest revenue films tend to follow the same pattern.

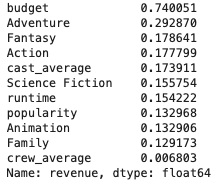
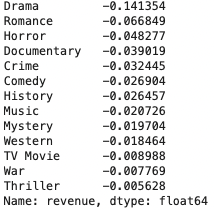


The above graphs show the distribution of movies in the data set with respect to genres (left) and the average revenue generated by a film of a certain genre (right). The highest grossing films tend to be Action, Adventure Fantasy, Sci-Fi, Family and Animation. Surprisingly highest grossing genres do not make up the majority of the data set, rather genres like Drama and Comedy do.

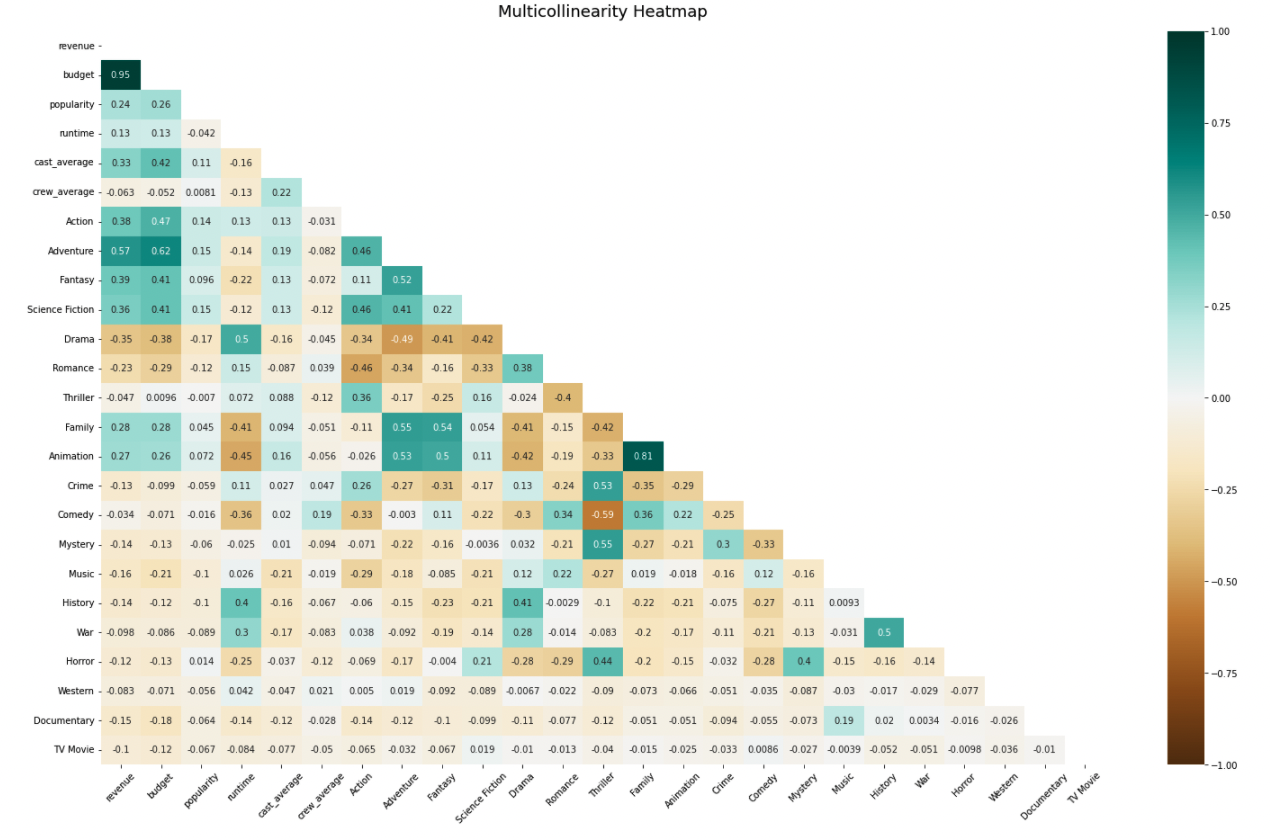


This plot illustrates the relationship between runtime and film revenue. Interestingly, even though the mean runtime across all the films in the data set is around 1.75 hours the above plot shows that the ‘sweet spot’ with regards to runtime is around 3.1 hours.

**3.1.5 Correlation Analysis and Multicollinearity Map**

One of the goals in this project is to accurately predict box office revenue for an upcoming film. Naturally this team analyzed the correlation coefficients between each of the attributes in the data set and the target (revenue) in order to determine feature importance. The above images show the results of that analysis. On the left, the attributes that are positively correlated with ‘revenue’ can be seen and on the right the ones that are negatively correlated with ‘revenue’ can be seen -- both ordered by strength of correlation. As can be expected, budget is most highly positively correlated with revenue, followed by several genres, cast’s average, runtime, and the film popularity. In terms of negative correlation, it was surprising to see that most of the genres -- with the exception of Fantasy, Action, Sci-Fi, Animation and Family -- are negatively correlated with ‘revenue’.



The heat map shown above describes the correlation (both positive and negative) between all the different attributes in the data set -- excluding production companies, actors and crew members for obvious reasons. Multicollinearity analysis again shows how correlated budget is with revenue because all attributes that are heavily correlated with revenue are also heavily correlated with budget. Interestingly, it also shows that films that fall under the genres Drama, Crime, History and War tend to have a higher runtime while films that fall under genres like Family, Animation, Comedy, Horror there tend to be shorter. In addition, it can be seen that several genres are highly positively correlated with each other (i.e. if a film falls under one genre, it is likely it will fall under the other) such as Family and Adventure, Family and Fantasy, and Thriller and Mystery/Crime. On the other hand some genres are negatively correlated with each other like Comedy and Thriller.

**4 Models**

**4.1 Model Selection**

The two regression models tested for this project were a Random Forest Regressor (Sklearn) and a Gradient Boosted Regressor (XGBoost). These models utilise two ensemble methods, namely, bagging and boosting. Squared Error was used as the objective function for both models. Our final dataset consisted of approximately 9600 records, so the team decided to use a 90-10 split for test and train data. In other words, 90% of the processed dataset was used for training and the remaining 10% was used to test our models.

**4.2 Model Tuning**

**4.2.1 Normalization**

We decided to log transform the budget and revenue (target) variables because they were highly skewed. This is likely because the amount of money spent on films (budget) and the amount generated from films (revenue) has risen exponentially over the last century or so.

**4.2.2 Hyperparameter Tuning**

We decided on using a hyperparameter grid search with 5-fold cross validation (GridSearchCV from Sklearn) to tune our model parameters. The aspect of cross validation allowed us to gauge the performance of our models on unseen data. With regards to the extreme gradient boosted regressor, we tuned the following parameters: max\_depth, n\_estimators, colsample\_bytree, min\_child\_weight, gamma and subsample. For the random forest regressor we tuned max\_features, max\_depth, n\_estimators, and min\_samples\_leaf. For each parameter in our parameter grid we tested a range of 3 to 5 values. This amounted to around 1800 parameter combinations for each model. Adding the element of cross validation, each of those parameter combinations was fitted 5 times summing to over 8000 unique fits per model. This took around 5 days to run on our local machines.

**4.2.3 Feature Selection**

After tuning our model parameters our focus shifted to feature selection. We used recursive feature extraction with cross validation (RFECV from Sklearn) to produce an optimal set of features based on validation scores and feature importance (a measure of correlation to the target variable). Of the 1312 features we began with, 551 were selected to be used in training and testing. The selected features consisted of the budget, release date, popularity, runtime, cast and crew average popularities, 308 actors, 111 crew members, 50 production companies, 19 genres, and 50 original languages. The features left out consisted mostly of extraneous actors and crew members. This gave us a better idea of which features we should look to expand on in future analysis. For instance, we know that the top 500 cast and crew members seemed to account for the ones most influential to film success, but in the future we might look to expand on the production companies we selected.

**4 Evaluations**

**4.1 Scoring**

For scoring, we decided to go with an R-Squared indication of success. This scoring metric essentially illustrates how well the model fitted the data to the regression line. In other words, it indicates how well the model explains the variation in the data.

**4.2 Predictive Capacity**

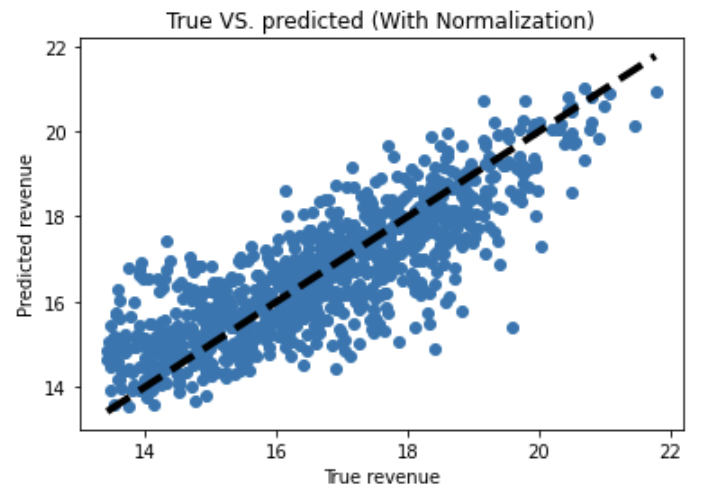
This project has been successful at finding and extracting trends in the dataset. The best performing model was the extreme gradient boosted regressor, producing an R-squared score of 0.694. As shown, there have been many trends found for the model.

This work can be evaluated largely by the accuracy of the prediction models and by utilizing the patterns evaluated from past data. If data up to a certain point in time in the past is used, and treats the more recent past as the future for the prediction model, then the predictions can be compared with the actual results that occurred (the baseline models), and thus determine the accuracy and success of the models. If these models align properly with the real-world occurrences (with a slight margin of error if need-be), then the models can be deemed successful and thus applied towards various films’ future predictions.

In order to judge the true success of this project, various metrics need to be used to accurately measure the validity of the results. These will largely be **accuracy** and **error**, as both of which take into account actual occurred data points and compare them against those predicted, and will produce a value that can be used to determine success. Time-centered metrics (i.e., latency) will not be used, as those are typically used in studies that center around performance rather than data analysis.

The baseline methods to compare with, and thus determine success, will be past data of films treated as upcoming occurrences, to then run the models and algorithms against and conclude whether or not they are valid.

**4.2 Notes on Tuning**

****

The normalization of budget and the target variable improved the performance of the model by around 11%. The hyperparameter tuning increased model performance by around 21%. The feature selection improved the model by around 5%. The above plot describes the observed versus predicted data plotted with respect to the regression line. It is evident that there are outliers present in the data. These movies that overperformed/underperformed with respect to our predictions can likely be attributed to factors not considered in this project. For instance, we had no way of gauging the publicity surrounding movies pre-release or whether or not the movie was a part of a larger, well/poor performing series.

**5 Future Work**

The future work for this project will be largely centered around increasing the number of parameters. This team was limited in the number of parameters that could be used in order to train/run the model within the timeframe of this project. So being able to use more variables from the original dataset could improve the R-Squared score. Also this team could improve the variables. Some of the variables that were decided upon were limited to only the highest performing of their class. For instance, the original number of actors and actresses were limited from the original very high amount to only 500. Also only 25 production companies were used and the thought process was because the top 25 production companies make up such a large majority of the data that the influence of the other companies would not impact the accuracy of the model very much.

**5.1 Evaluations**

We were able to evaluate a few of the top correlated variables to overall gross revenue of a film. As you may assume a major part of a movie's success is the actors and actresses they bring on. Securing big blockbuster talent is vital for a lot of films in order to get an initial reaction from customers to get them in the door. The top ten performers correlated to a high grossing revenue were interestingly all male actors such as, Frank Welker, Samuel L. Jackson, Liam Neeson, Robert De Niro, Bruce Willis, Nicolas Cage, Morgan Freeman, Willem Dafoe, John Goodman, and Steve Buscemi. Another highly correlated variable to the success of film were the crew members such as directors and writers, the top of their class were Peter Hyams, Juan Peralta, Marshall Winn, Shelley Roden, and J. Michael Muro. Finally, as you could guess, the production company to produce a film was really important to a movie's gross revenue. Production companies with a long track record and an overall high prestige bring a lot of value to a movie. The top five production companies wereWarner Bros. Pictures, Universal Pictures, Paramount, 20th Century Fox, and Columbia Pictures.

**6 Conclusion**

Overall we were able to achieve the parameters of the project we set out for when we started this project. As mentioned before we measured our accuracy with the R^2 score and got it to 0.694 which from what we could tell is in the same ballpark score as similar projects. However there were definitely some things we could have done differently to improve overall. Firstly we think our model could have really benefited from adjusting the data we used, to account for the rise in inflation. The revenues and budgets of older films are recorded as they were without edit, so that creates a potential skew in our model that we could negate by adjusting the prices of those variables to the current value of the dollar. Along with the inflation there is another potential skew in the data from the fact that movies today have drastically higher budgets as a result of the market just being so much bigger. As a result, variables we use, such as production companies, could lose accuracy just due to the fact that the times have changed and companies are willing to spend much more than they used to. For example, Paramount picture studios is in our top 3 production companies, and it has been around since 1912, and therefore the value they can add to a movie's revenue is much greater now than it was 100 years ago. To build on that we could have broken up movies into separate eras to account for this skew. For instance breaking up our model to only movies from 2010-2020 could make the model more accurate for movies of that era.

**6.1 Limitations**

A major limiting factor in this project for us was not being able to tinker and tune every little attribute and parameter in our models as freely as one would have hoped for. A model normally would take days to finish, as a result we really had to be deliberate in the changes we made to the model. With more time and computing power we would have been able to make those small changes mentioned above to really fine tune the model and achieve a higher R^2 score.

**7 Milestones\***

∗The decisions on the structure and time allowances of each subtask were based upon the team’s past experience with tasks of a similar workload, and portioned out accordingly

The following is the final schedule for the completion of this project, laid out in such a way as to guarantee a smooth and efficient workflow and an appropriate finish time:

**Week 1 (10/4):** ~~Gather all necessary information, data and API permissions necessary to analyze. Store into the database, begin to write queries to analyze.~~

**Week 2 (10/11):** *~~Finish query analysis~~*~~, begin normalization of the data.~~

**Week 3 (10/18):** *~~Finish normalization of the data~~*~~, begin pattern analysis (using queries written before). Log patterns and determine where else similar patterns may lie.~~

**Week 4 (10/25):** *~~Finish pattern analysis~~*

**Week 5 (11/1):** *~~Begin drafting the report.~~*

**Week 6 (11/8):** *~~Normalization and analysis finished, Begin construction of models for future prediction~~*

**Week 7 (11/15):** *~~Continuation of model construction, Begin writing of final report~~*

**Week 8 (11/22):** *~~Finished rough construction of models, Begin testing of models, Continue writing of final report~~*

**Week 9 (11/29):** *~~Finish final models & predictions, Finish rough draft of report, Begin final presentation~~*

**Week 10 (12/6):** *~~Finish final draft of report, Finish final presentation~~*

**9 Work Breakdown**

Below are all of the assignments worked on and the code written.

Connor Dixon:

* Assignments Worked On:
  + Course Project Proposal Report
  + Course Project Proposal Slides
  + Course Project Checkpoint Summary Slides
  + Course Project Checkpoint Report
  + Course Project Final Presentation Slides
  + Course Project Final Report
* Code Written:
  + models.py (tests a GaussianNB model against processed data)
  + new\_models.py (runs a GridSearchCV to determine best parameters for RandomForrestRegressor)

Kaylee Engelhardt:

* Course Project Announcement
* Course Project Proposal Report
* Course Project Proposal Slides
* Course Project Checkpoint Summary Slides
* Course Project Checkpoint Report
* Course Project Final Presentation Slides
* Course Project Final Report

Baci Brunet:

* Assignments Worked On:
  + Course Project Proposal Report
  + Course Project Proposal Slides
  + Course Project Checkpoint Summary Slides
  + Course Project Checkpoint Report
  + Course Project Final Presentation Slides
  + Course Project Final Report
* Code Written:
  + api\_call.ipynb
  + preprocessing.ipynb
  + plots.ipynb
  + hyperparameter\_tuning .ipynb
  + feature\_selection.ipynb
  + model.ipynb

Aidan Aarts:

* Course Project Announcement
* Course Project Proposal Report
* Course Project Proposal Slides
* Course Project Checkpoint Summary Slides
* Course Project Checkpoint Report
* Course Project Final Presentation Slides
* Course Project Final Report

**10 References**

1. Eon Smit and N. A. Pangarker. 2013. The determinants of box office performance in the film industry revisited. Research Article. South African Journal of Business Management, (September 2013). https://www.researchgate.net/publication/281730174\_The\_determinants\_of\_box\_office\_performance\_in\_the\_film\_industry\_revisited
2. Jia Xiao, Xin Li, et al. 2017. An inside look into the complexity of box-office revenue prediction in China. Research Article. SAGE Publications, (January 2017). https://journals.sagepub.com/doi/10.1177/1550147716684842
3. Sophia Gaenssle, Oliver Budzinski, and Daria Astakhova. 2019. Conquering the Box Office: Factors Influencing Success of International Movies in Russia. Research Article. Review of Network Economics, (October 2019) https://www.degruyter.com/document/doi/10.1515/rne-2019-0017/html