Tables of Contents

Introduction

As time goes on and viewers search for new entertainment, the film industry has attempted to adapt to expectations and create movies that keep customers returning to see more. Our goal for this analysis is to predict film revenues as well as determine whether a film will have "high revenue," meaning their revenue is five times higher than the average budget.

The dataset for the analysis is from a Kaggle's Box Office Prediction competition located at https://www.kaggle.com/c/tmdb-box-office-prediction. The training and test data files provided have over 7,000 rows (movies) total; 3,000 observations in the training set and 4,398 in the test set. The Kaggle competition hosts noted there are no duplicate movies although there may appear to be due to remakes, multiple films with the same title, etc. There are 22 columns in the datasets including budget, genres, overview, popularity, release date, cast, and crew. The training dataset includes a revenue column, which is the target dependent variable and is not included in the test data.

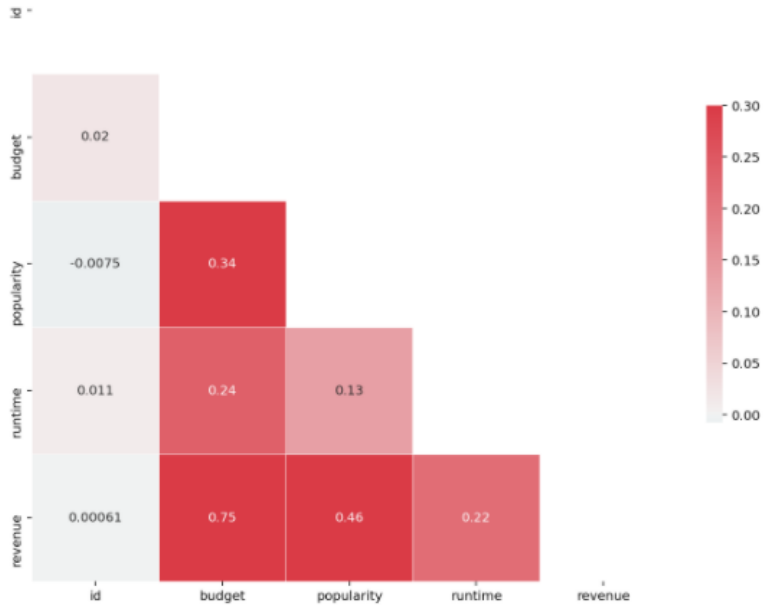
We begin our analysis by completing Exploratory Data Analysis (EDA) on the variables provided to visualize how each feature relates to one another and how they then compare to revenue.

EDA & Feature Engineering

Exploratory data analysis (EDA) is used to visualize data sets and highlight their main characteristics. By plotting graphs and creating visualization, we are able to explore the common features in film revenues. EDA won't necessarily predict which characteristics generate the majority of the film revenue, but it is a useful stepping stone to direct our future analysis. We want to ensure we're not making assumptions or final predictions solely based on the visualizations created.

Feature engineering is an important process that helps extract features from the dataset to improve the performance of machine learning algorithms. This process is crucial to create predictive models in the future.

We began with a correlation matrix of the numerical variables: revenue, runtime, popularity, budget, and id. Based on the plot below, we can see budget is highly correlated to revenue and popularity and runtime are moderately correlated to revenue. Those three variables appear to be significant in predicting revenue and will most likely be included in the predictive models. We will conduct further analysis to be sure how best to use them in the models.

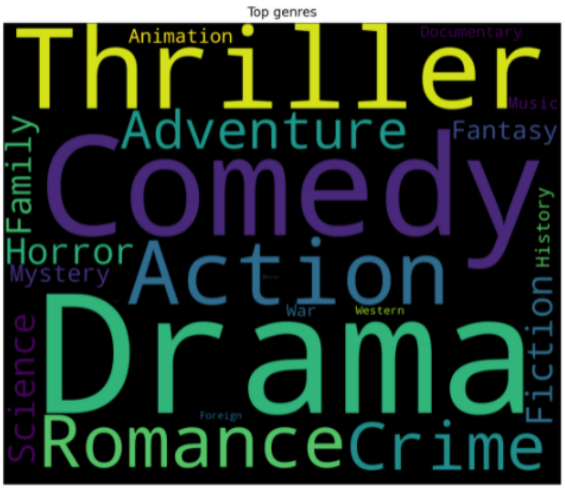


Belongs_To_Collection

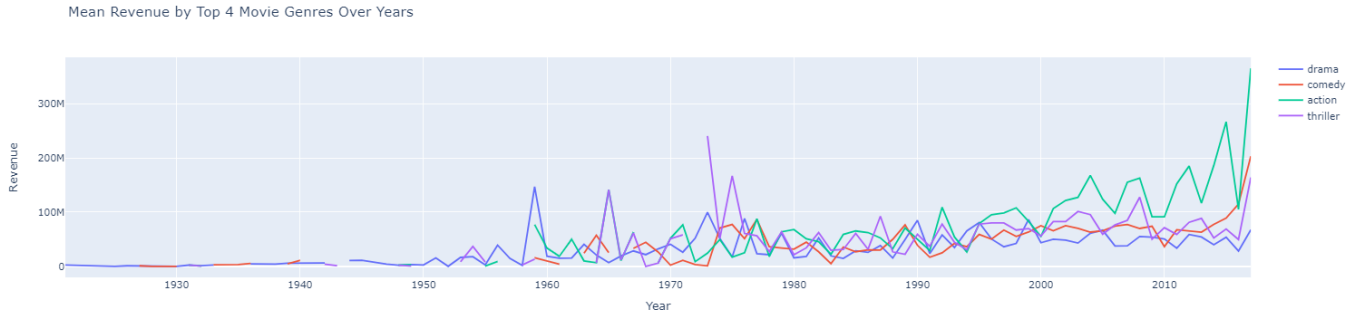
The belongs_to_collection column is in a dictionary format of the ID number, name, and marketing material website paths of the collections. This column needs to be parsed into useful columns, which contains the collection name each movie is a part of and if the movie is in a collection. While, only about 20% of the movies in the training set are in a collection, it stands to reason that if a movie is part of a collection it could impact box office revenue. The plot below shows films in a collection (orange) earn higher revenues.

Genres

The genres column is the ID number and name of all of the film genres a movie belongs to. Most have two to five genres with zero, six, and seven as outliers. The word cloud to the right shows the most common genres are drama, comedy, thriller, and action. We created separate columns for the 15 most common genres and a column with the number of genres in each film to be used in the analysis.

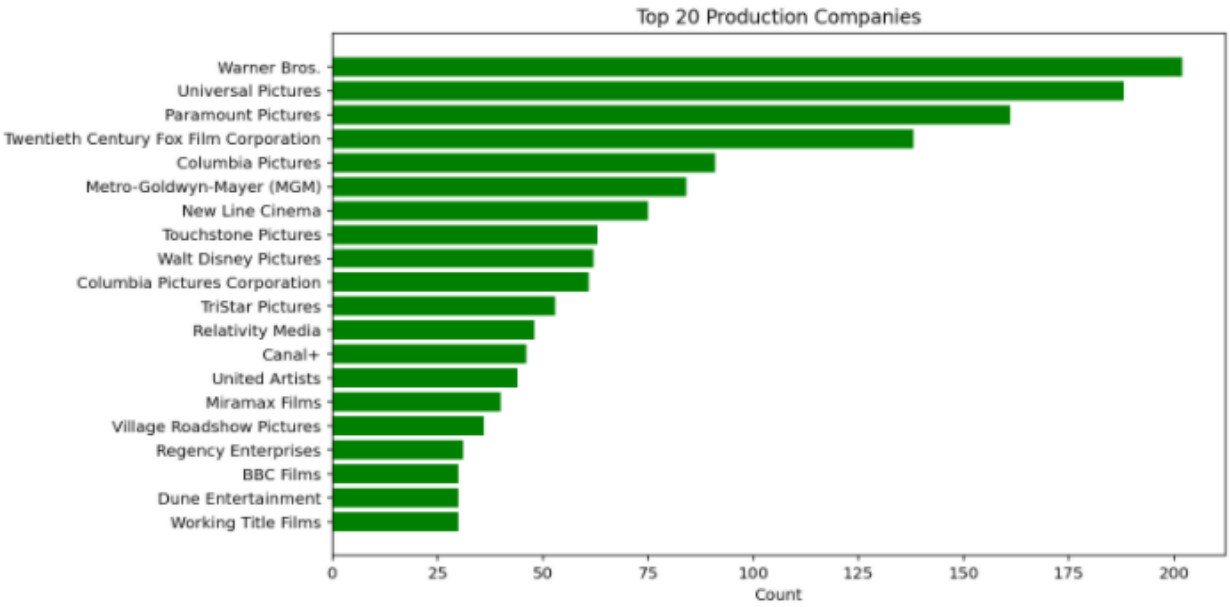


The plot below further reflects the relationship between genre and the target, revenue. Action films tend to earn higher revenue, which has increased in the past few decades (at least partially due to inflation).



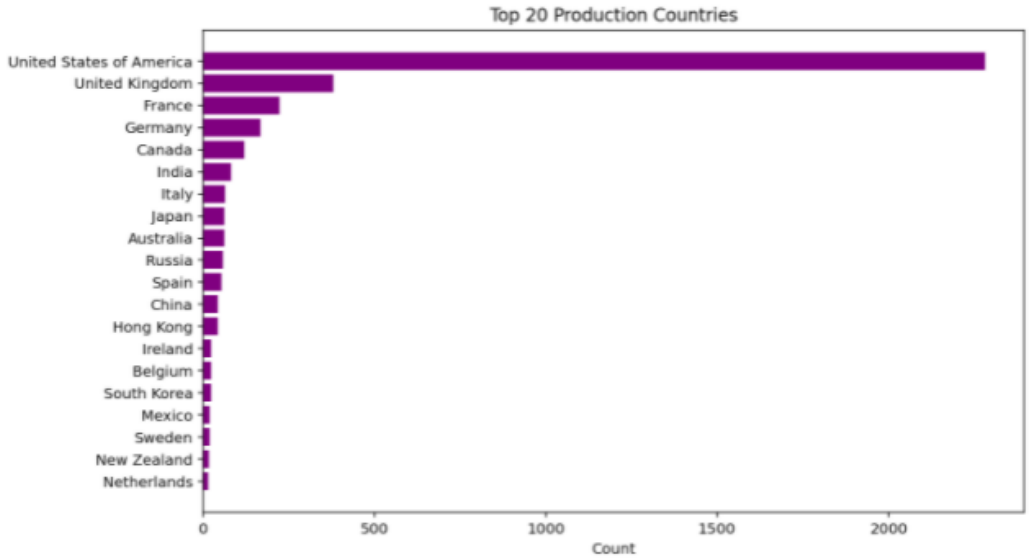
Production_Companies

The production_companies column has the name and ID number of all of the production companies involved in each film. The majority have one to nine production companies with a few that have 10 or more. Like several other competitors in this Kaggle competition, it is unclear if this information will be useful in the analysis, so we created binary columns for the 20 most common production companies and a column with the number of production companies for each film. The top 20 production companies are shown below. Warner Bros, Universal Pictures, Paramount Pictures, and Twentieth Century Fox Film Corporation represent a good portion of the films in this data set.



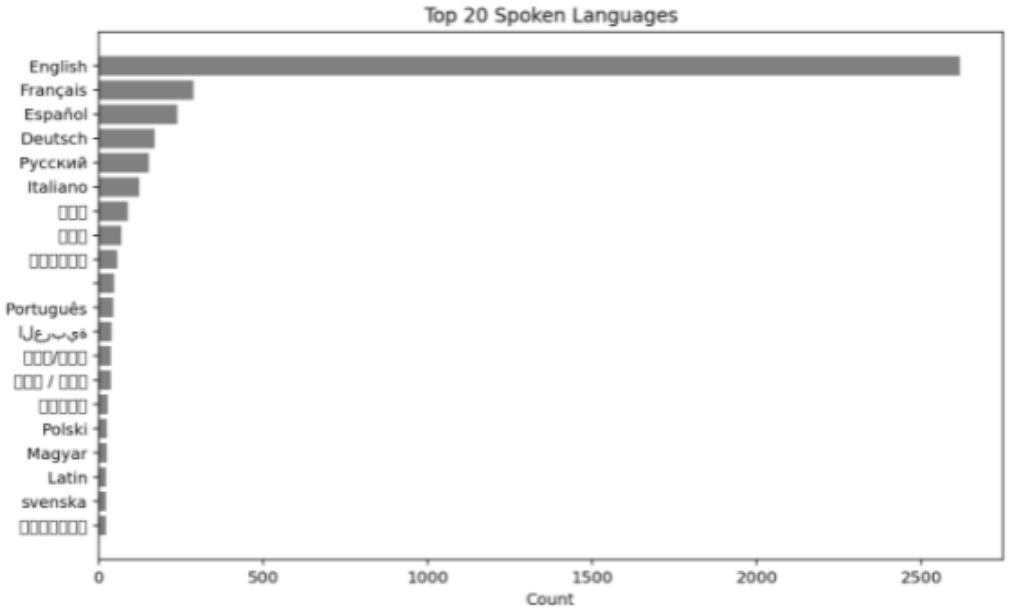
Production_Countries

The production_countries column shows the country name and abbreviation for all production countries involved in each film. A majority have just one production country, the United States, but some films have up to five. There are just four films with more than five production countries. Similarly to the production_companies, it is unclear if this information will be useful in the analysis, so we will create binary columns for the 10 most common production countries and a column for the number of production countries for each film. The top 20 production countries are shown below, which shows a high skew toward the United States.



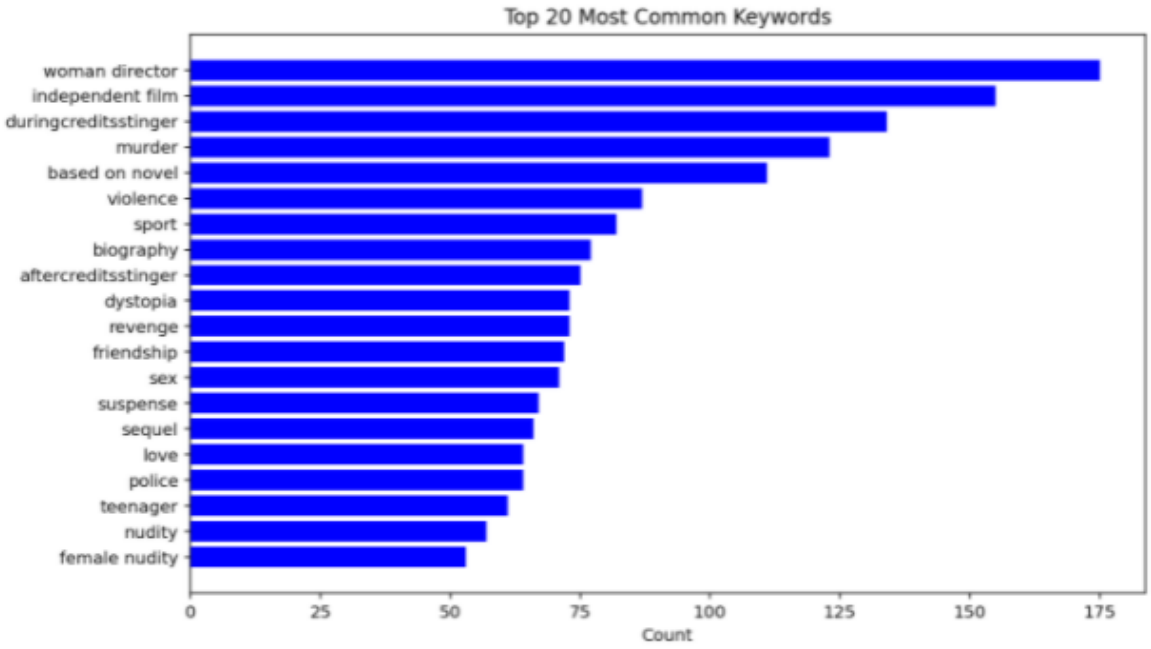
Spoken_Languages

The spoken_languages column has the abbreviation and language name of the languages spoken in each film. As shown in the barplot below, English is by far the most commonly spoken language in the films in this dataset. With such little variability in the column, it does not appear it would be a useful factor in the analysis. However, to test that theory, we will follow the same pattern as the columns above and create binary columns for the top 10 most common languages, as well as a column with the number of languages spoken in each film.



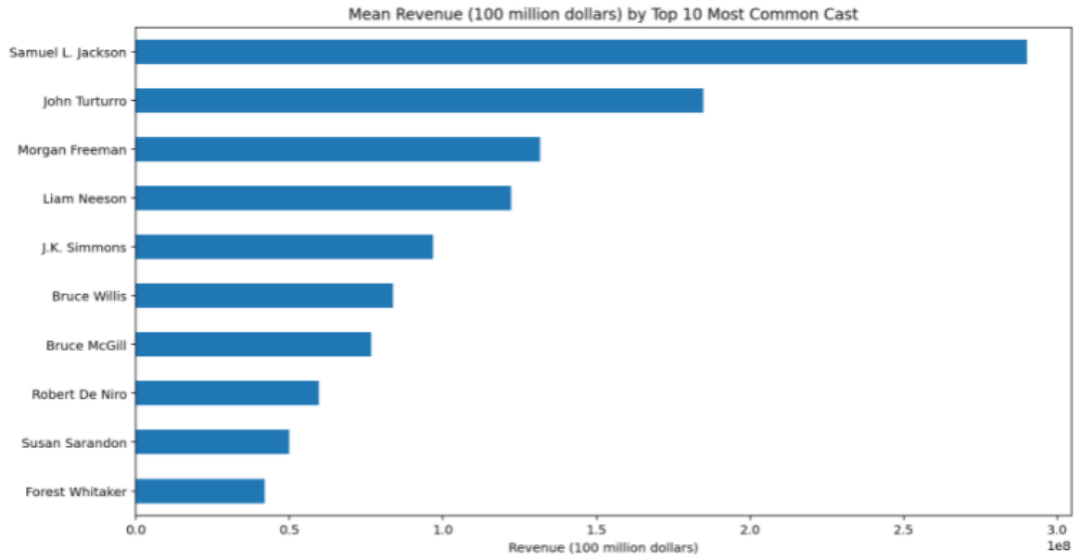
Keywords

As you would imagine, there are a lot of keywords used in film reviews and film marketing, such as woman director, independent film, and murder, as you can see in the word cloud to the right and barplot below. To make better use of this information, we created a column with the number of keywords used for each film and corresponding binary columns for the 30 most common keywords.



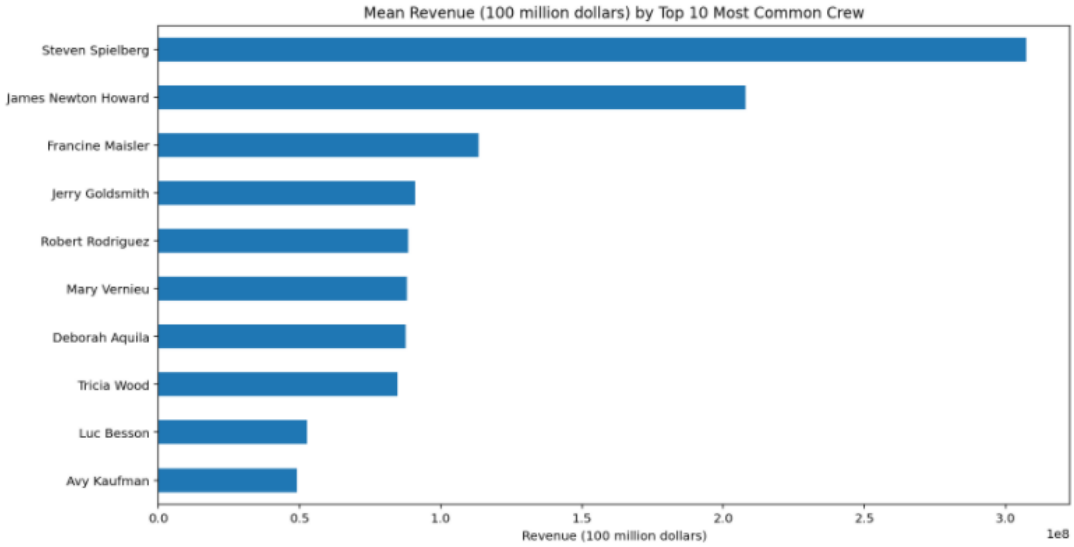
Cast

There are a lot of cast members per film, typically between 9-18, and as suggested in a study by Bhave, et al, the cast can be very influential on the box office revenue (2015). The data includes actor name, gender, and character/type. As documented on the competition site, gender is encoded as 0 = unspecified, 1 = female, 2 = male. In similar fashion to the EDA above, we created a column with the number of cast members, and associated binary columns for gender and the 15 most common actors and characters. The mean revenue by top 10 most common cast members is seen below. Samuel L. Jackson, John Turturro, Morgan Freeman and Liam Neeson are among the Top 4.



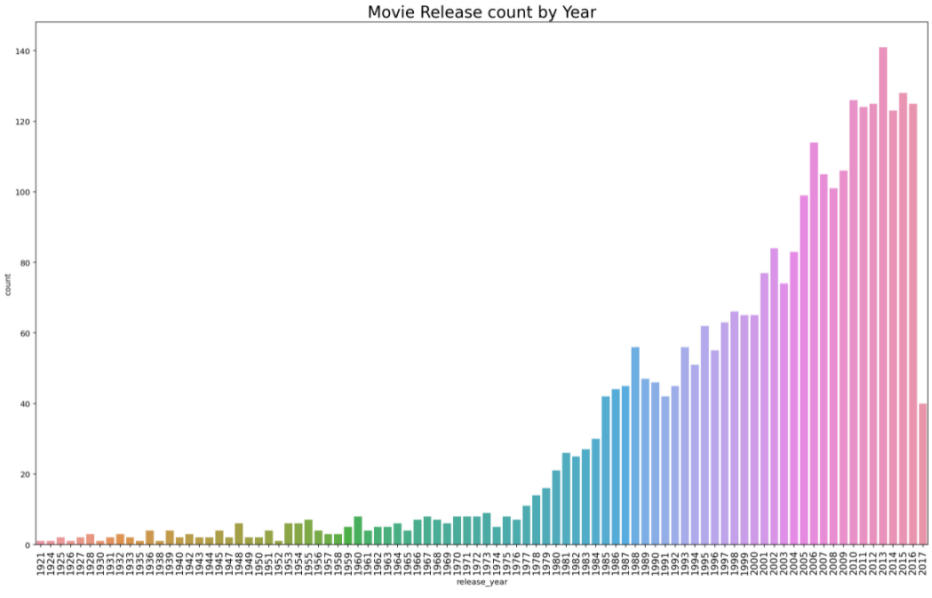
Crew

Bhave, et al, also suggested a film's crew was influential in box office revenue (2015). The data in this column includes crew name, gender, department and job/role. As we have done previously, we will create a column with the number of crew members, and binary columns for gender and the 15 most common crew members, departments, and jobs. The mean revenue by top 10 most common crew members is below. The Top 3 in this data set are Steven Spielberg, James Newton Howard, and Francine Maisler.



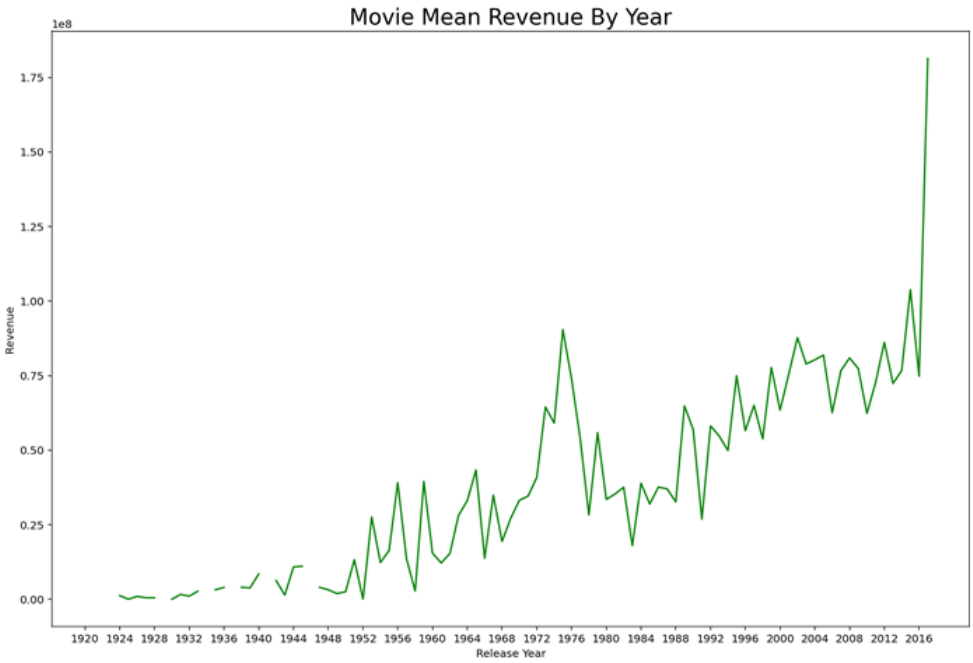
Release Date

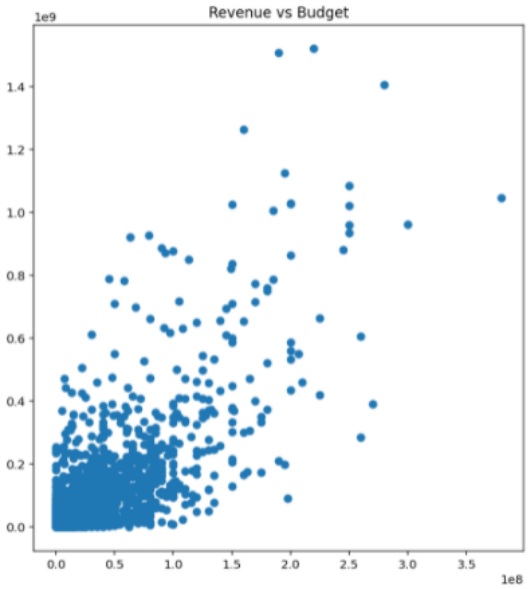
Only the last two digits of the release year were provided, so we first had to format the column to view it correctly. The distribution of the number of films released by year is shown below. You can see below that the years are skewed left, so the 2000's and 2010's represent a majority of the film records.

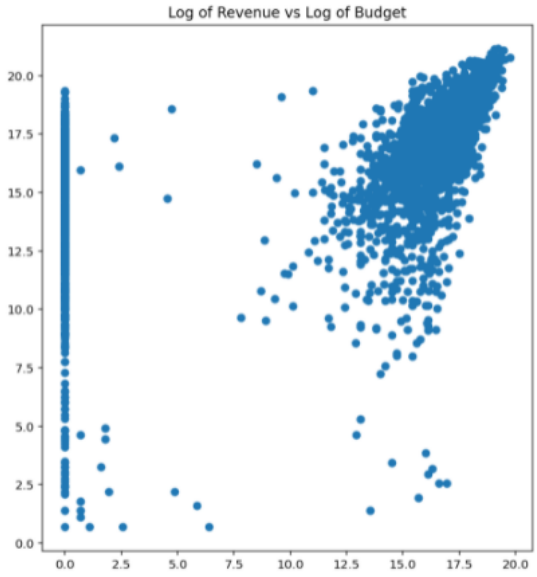


Revenue & Budget

The distribution of revenue and budget is also very skewed to the left, so we also took the log of both variables to better manage them. However, using either the raw data or the log values, revenue and budget are highly correlated - the higher the budget, the higher the revenue.

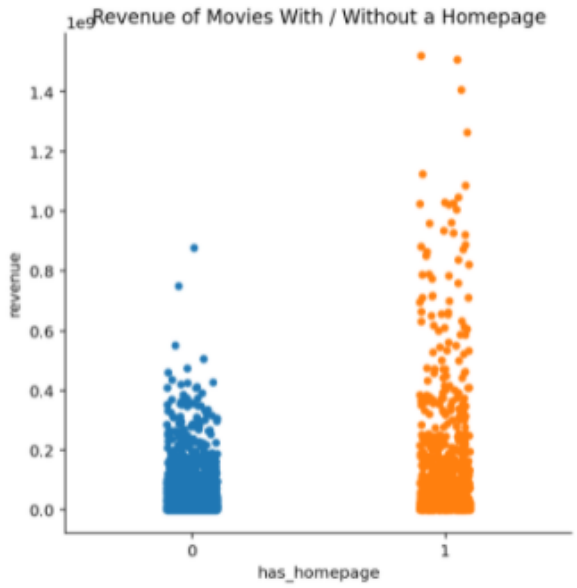






Homepage

The homepage column has a lot of unique variables, so we categorized them into either having a homepage or not. As shown below, films with a homepage (orange) earned higher revenue than those without a homepage (blue).

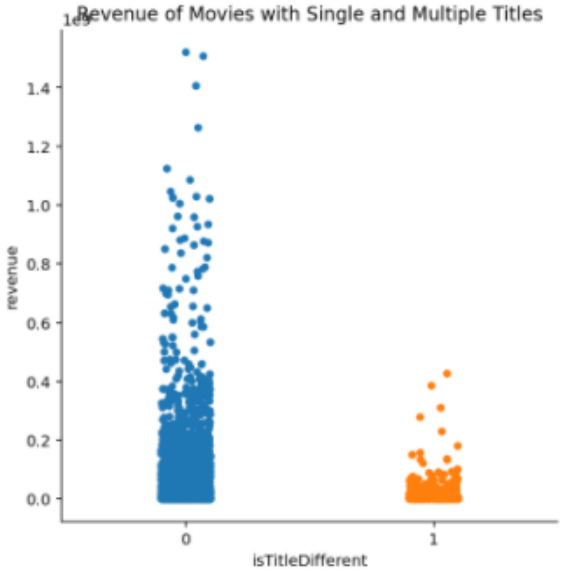


Tagline

In similar fashion to the homepage column, the tagline column has many unique values, therefore the column was categorized into those with or without a tagline. As shown below, films without a tagline (blue) earned higher revenue than those with a tagline (orange).

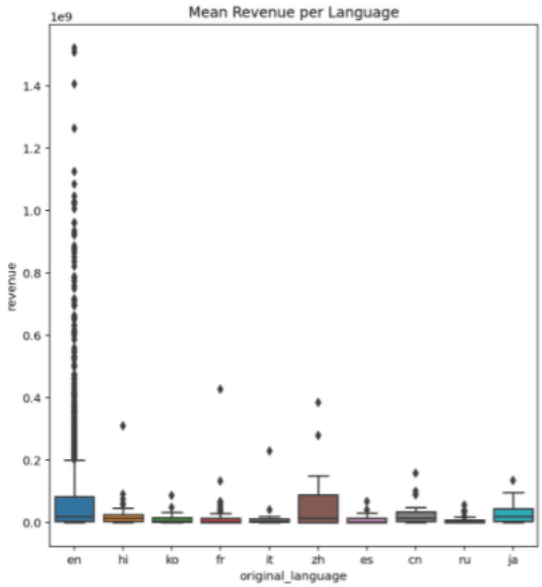
Original_Title and Title

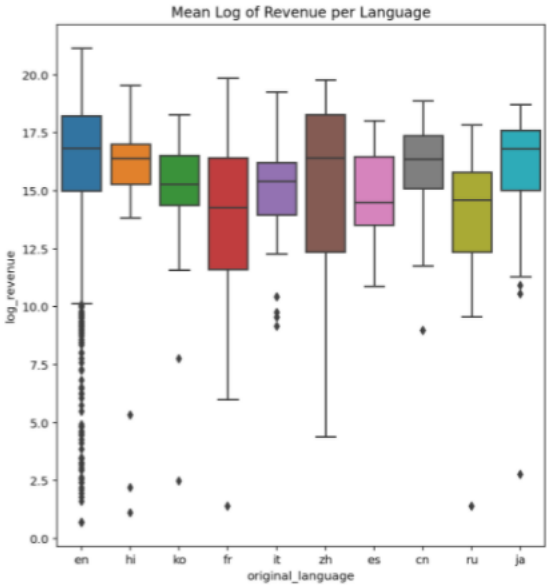
Some films have more than one title depending on where they are released. To best analyze the data, we looked at revenue by films with a single title compared to those with multiple titles. Films with a single title (blue) earned higher revenue than filmes with multiple titles (orange).



Original_Language

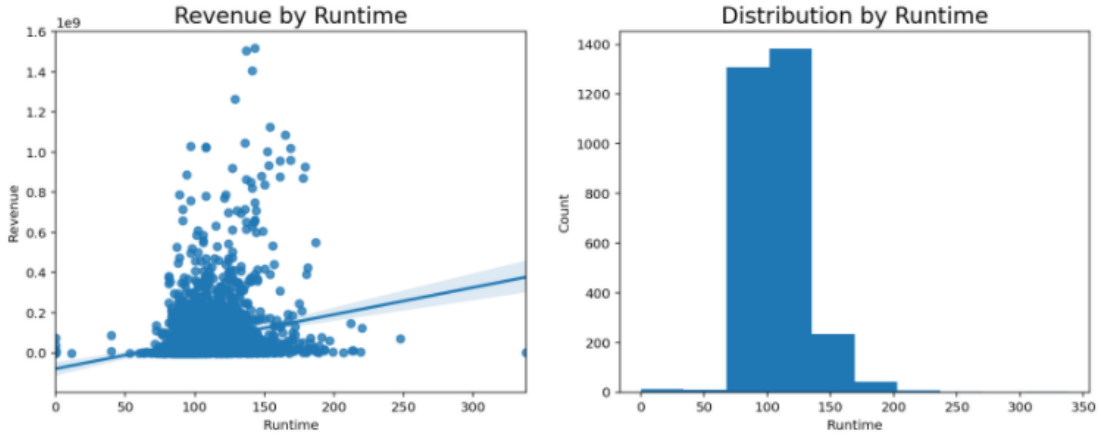
The majority of the films are in English and as shown below, earn the highest revenue. However, when compared to the log of revenue, the distribution appears much less skewed.





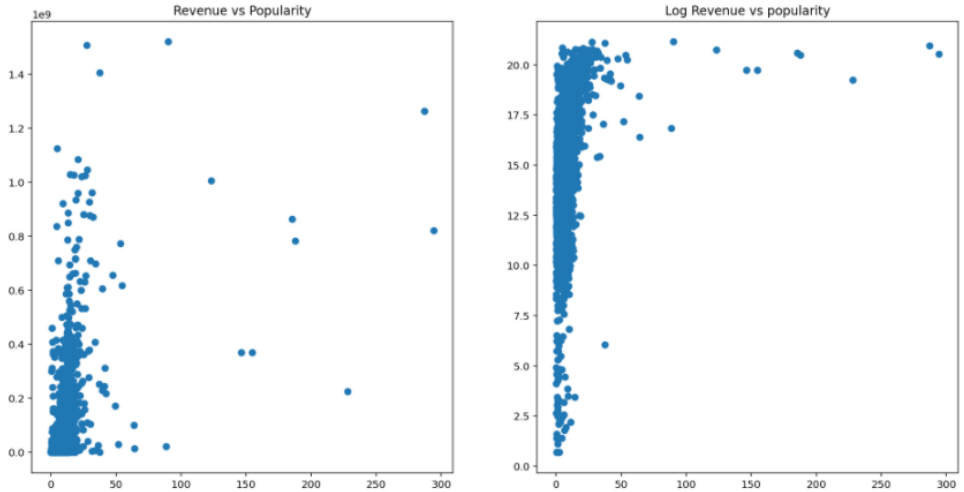
Runtime

While not entirely clear, it appears runtime may impact film box office revenue. As in our initial correlation analysis, runtime is moderately correlated with revenue, as shown below. Runtime is also skewed to the right as the majority of films are less than 200 minutes long.



Popularity

As with runtime, popularity also seems to be correlated with revenue. It is unclear how the competition host calculated the popularity of a film, but it is possible it is based on ratings or reviews of each film.



Regression Predictive Models

Regression analysis is a type of predictive modeling that is used to predict a continuous variable based on one or more independent variables. Regression models help explore what relationship is significant between dependent and independent variables. What this means is we can make predictions of which independent variables will have a higher impact on our dependent variable. The two regression models we chose to analyze are linear regression and random forest regression.

To begin running the regression models, we first preprocessed the data. Because we are dealing with a moderately sized dataset, we created a reduce_mem_usage function to reduce the data frame size. By using this function, we were able to reduce the memory usage for the train dataset by 83.3% and 84.3% for the test dataset. This helps our code execute faster since we are not storing all results in memory..

In order to run the code efficiently, we cleaned the datasets by identifying columns with missing values. We found that 97% of the meanBudgetByYear and meanRevenueByYear variables included missing values. We also retrieved the categorical columns in both the train and test datasets. For these categorical columns, we set the missing values to “na”. Lastly, we subsetted our training dataset to only include categorical variables and generated a new dataframe ‘x’. We then used this ‘x’ dataframe to convert each value in each column to a number by using the label encoding technique for categorical variables.
To continue with our modeling, Python offers a tool in scikit-learn called Pipelines that helps with automating the work required for machine learning. This tool allows us to evaluate the models we are using by cross-validating and providing a score. The purpose of this tool is to make sure that the steps in the pipeline are restricted to the training dataset for each fold of cross validation. We then use grid search cross validation ‘GridSearchCV’ in order to implement a fit and score method. GridSearchCV helps us find the best parameters from the set we give. These tools will allow us to efficiently find the best model.

Linear Regression

Linear regression is used to establish a relationship between our dependent variable and one or more independent variables. This is implemented using a regression line - line of best fit. This method assumes the predictor variables are normally distributed with constant variance. It is ideal to minimize multicollinearity with other predictor variables. One thing to note is, this model also does not handle outliers well and they can heavily impact the regression line and associated forecasted values.

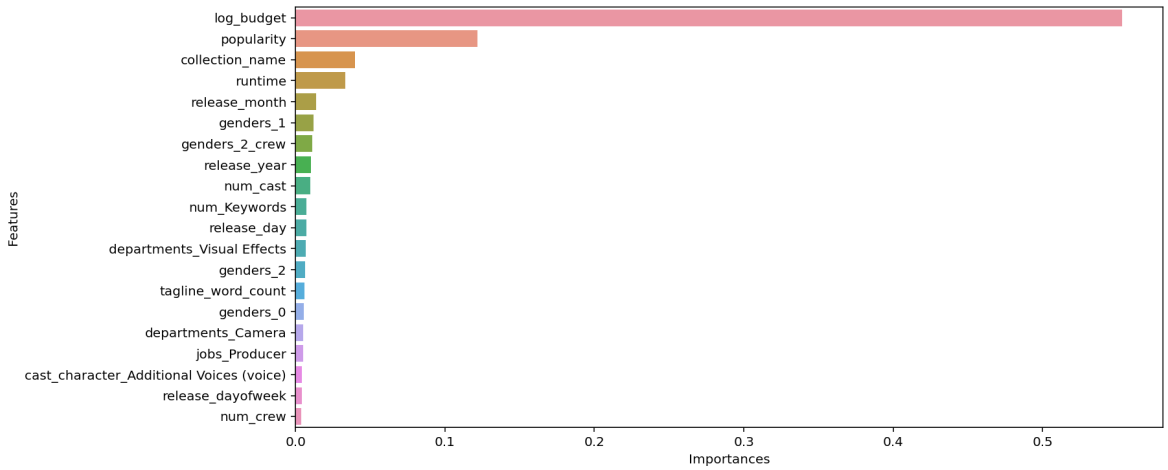
In our application, our model was fitting 10 folds for each of the 247 candidates, totalling 2,470 fits. Our cross-validation accuracy score for this model was -3.19975600321177e+23, which translates to having the mean square error of approximately 3.1998e+23.

Random Forest Regression

Random forest regression is a supervised machine learning algorithm that uses an ensemble technique. It is used to fit a number of classifying decision trees on sub-samples of the dataset. It then uses an averaging method to improve the accuracy of the prediction. It also helps control over-fitting issues when running the model. This regression method should be used when there aren't any linear trends in the training dataset. Also, a study by Lee, et al, found that decision tree models, such as random forest, result in better box office revenue predictions than other model methods, such as boosting or bagging (2020).

For these datasets, we are setting the number of trees in the forest to 200. As mentioned earlier, this method uses the GridSearchCV to find the optimal values for us to use for the hyperparameters of our model. After running the model, we calculated the accuracy score and found that the random forest regressor is about 96% accurate for the train dataset and 62% for the test dataset.

After running the random forest regression, we also looked at the features importances in the data frames. The bar graph of features importances below shows the random forest assigned log_budget to be the highest in importance, which means that the budget has the most impact on a film's revenues by far. Popularity and collection_name are the second and third most important features, respectively.



By looking at both linear and random forest regression predictive models, we can see that the random forest regression model has a higher accuracy. This is a more efficient method as the data has a nonlinear relationship.

Classification Predictive Models

Classification predictive modeling is slightly different from regression predictive modeling. As we saw in our previous regression predictive models, it approximates the independent variables to numerical or continuous dependent variables. Classification predictive modeling, on the other hand, estimates the independent variables to discrete or categorical dependent variables. The two classification predictive models we chose to analyze are logistic and random forest classification.

To begin running the classification predictive models, we used the reduce_mem_usage function to reduce the data frame size. Because of this function, we were able to reduce the memory usage for the train dataset by 9% and 9.4% for the test dataset.

In order to run the classification predictive model, we have to create a classifier variable with the value 0 or 1. We used Pareto's Law of 80/20 to set our threshold, meaning approximately 20% of the movies were above the threshold for high revenue, and 80% below. This classified variable we created for revenue is called ‘classified_revenue’.

Logistic Regression Model

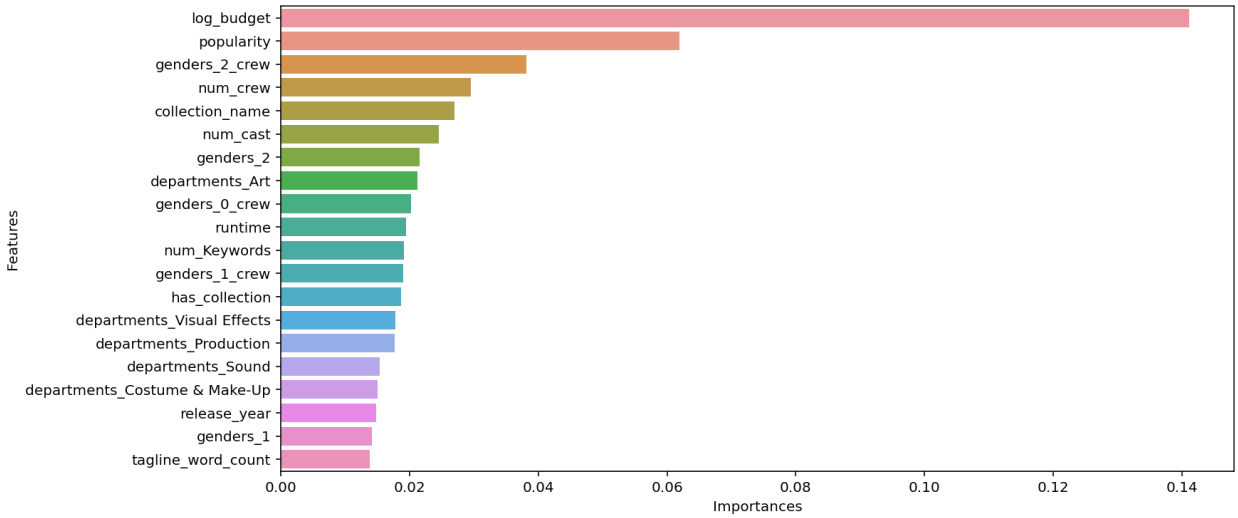
Logistic regression model is a type of classification algorithm as its target variable (or the dependent variable) is categorical. This is a well known method as it is used with binary output either in class or 0 or 1. Since we have created a classifier variable ‘classified_revenue’ with value 0 or 1 based on the threshold we previously set, we will be using that as our target variable.

After running this logistic classifier method, we used the predictor scores to compute the area under the receiver operating characteristic curve (ROC AUC). We found that the accuracy of this method is about 88.33%.

Random Forest Classifier

Random forest classifier method has a similar process as the random forest regression. Again, the difference is that our target variable is now a binary variable instead of continuous. After running this classifier method, we found that the accuracy of this method is 100% for the train dataset and about 90.83% for the test dataset.

Again, we reviewed the features importances, which are shown in the bar graph below. The graph shows the random forest classifier assigned log_budget to be the highest in importance, which means that the budget has the most impact on a film’s revenues. It is then followed by popularity and genders_2_crew, which is slightly different from the random forest regression features importance we previously ran.



In this case, we can see that the random forest classifier is not significantly more accurate than the logistic regression. The score differs only by approximately 2%.

Lessons Learned

Something we noticed while running the analysis is running into overfitting issues. This means the model that we ran has learned the training data too well. This could lead to inaccuracy or in sufficient prediction of the revenue on the test data. There are multiple reasons why our models are overfitted, which include having too many features in our training dataset. Having too many features can increase a model's complexity and redundancy of features that are unrelated to the prediction. In addition, we have a relatively small dataset, which could also contribute to the overfitting models. In the future, we could potentially use a linear regression model to select important features to train our model.

As this was a Kaggle competition, the end goal was to submit a file containing our predictions into Kaggle and receive a score. The initial plan was to submit the predictions made by the Random Forest Classifier, as this model had the most accurate predictions of revenue in the test dataset. However, there were some errors that arose while trying to create a submission file, due to differing lengths of indexes. Due to several external constraints, we were unable to fix this error, and could not submit a file to Kaggle in time.

Sources

Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015). Role of different factors in predicting movie success. *2015 International Conference on Pervasive Computing (ICPC)* (pp. 1-4). IEEE. https://ieeexplore.ieee.org/abstract/document/7087152?casa_token=gN6aYMDu5cgAAAAA:C1rOG6i8RNWVqjPGh1h3LiUvze2cToVNd_zXbjO7Z3y0pD34IciYEzL7ga1e_7ManugfMP1bhg

Lee, S., Bikash, K. C., & Choeh, J. Y. (2020). Comparing performance of ensemble methods in predicting movie box office revenue. *Heliyon*, *6*(6), e04260. https://www.sciencedirect.com/science/article/pii/S240584402031104X