Movie Insights

Greg Bodik, Marvin Yang, Kevin Zhou

1. Overview

The goal for this project was to investigate movie ratings and general measures of public opinion, with the hope of relating them to measures of films' economic performance and box-office success. In doing so, our main objective was to learn and implement the steps of a data-science oriented workflow, first collecting and exploring the data, deriving its most meaningful features (those that accounted for most of its variance), analyzing them, and then drawing conclusions from the resulting interactions. We used several different machine learning algorithms to answer both classification and regression questions of interest. Each of these aspects is detailed in the sections that follow.

2. Methods and Results

2.1 Data collection and motivations

The data in use comes from a Kaggle submission containing information on 5,043 films as originally obtained from the internet movie database (IMDb). The dataset had metadata for 24 different aspects of the films involved, and included movies from around the world. As we had originally wanted to do a project based on rankings, we thought that these data were interesting in that they involved readily understandable aspects of movie production (an already mainstream aspect of life) that we could use to gain a deeper insight into exactly the characteristics that differentiate successful from unsuccessful films, and make some films last while others fade away.

2.2 Data exploration

When first exploring the data, we noticed a division in the features between those based on the movies themselves (what actors were in the film, who directed it, its title) and information about how the film was received i.e. how many people left IMDb ratings, how many Facebook likes the actors in the cast recieved, what the movie's budget was, and how much money it ended up grossing. In particular we

thought it would be interesting to try to relate features together. When looking at metrics involving money, we quickly noticed a disparity between various global currencies and decided to omit those 1,236 movies that were not from the United States from consideration. Although we probably could have converted the revenue and budget amounts of the foriegn films to their USD equivalents, we chose not to because we did not want to introduce the variation associated with movie viewership in different countries, social media use, the prevalence of IMDb (which to our knowledge is mostly of use in North America) etc. An example of the effect of removing foreign films on movie budgets is seen in **figure 1**.

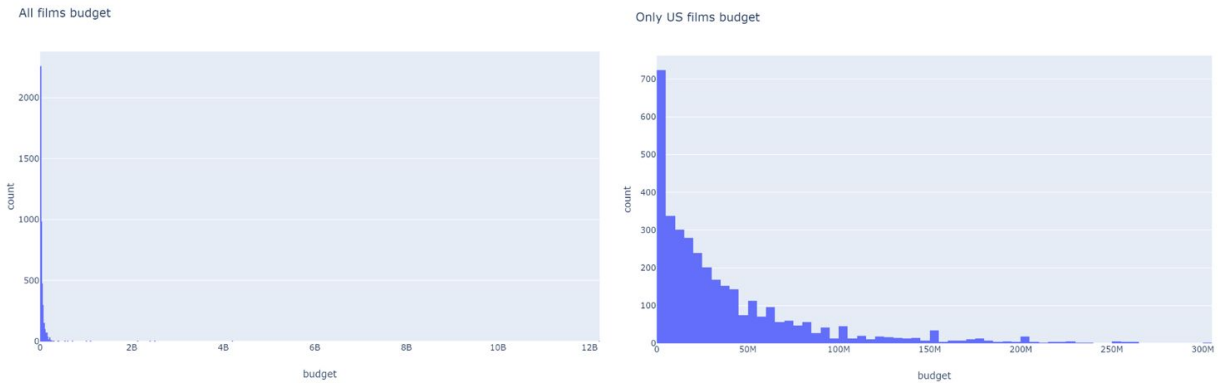

Figure 1: Distribution of movie budgets before and after foreign films were removed from consideration.

As will become important later, in the data exploration stage, we found that the distribution of the films' title years was severely skewed left, with some films as old as 1910 represented alongside the majority of more recently released ones. In our initial explorations, we identified two trends that we wanted to study further and that would drive our analyses. A movie's content rating denotes the audience for which it is deemed suitable. We noticed that films that appealed to a wider audience (such as movies rated G) usually had higher gross revenues than more narrowly targeted films (rated R for example). This is seen visually in **figure 2.** Secondly, when looking at the distribution of overall revenues, we noticed that it was severely skewed to the right, such that on average, films could be expected to gross fairly consistent amounts, save for a few extremely high outliers. To further probe this disparity, we wanted to see how revenue was positioned among films made in different years. Interestingly, median revenue of films from each particular year was relatively constant (**figure 3**). Given the extremely skewed nature of the

overall revenue data, this would mean that each year must have only a few outliers that contribute to the shape of the overall distribution, but that there is consistency from year to year. The findings noted above, in addition to our other explorations of the data led us to strive to answer the following questions in our project:

1. What characteristics are common to highly grossing films, and how can we predict how much a film will gross based on them?
2. How much does the content of a film impact how it is received? Or conversely, does there exist a pattern in movie engagement from viewers such that we can predict the content rating of a movie?

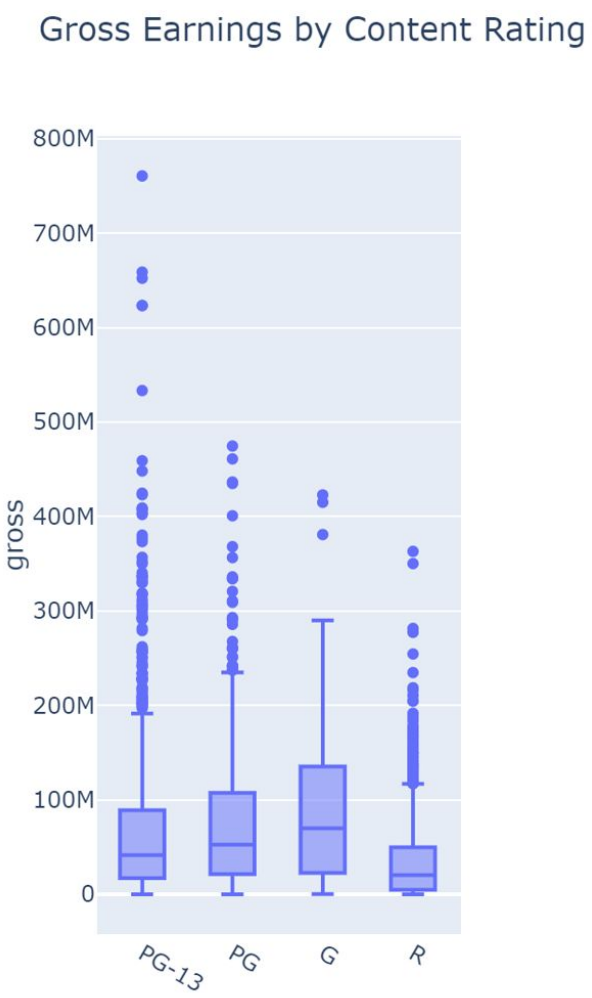

Figure 2: Films appealing to wider audiences had greater gross earnings.

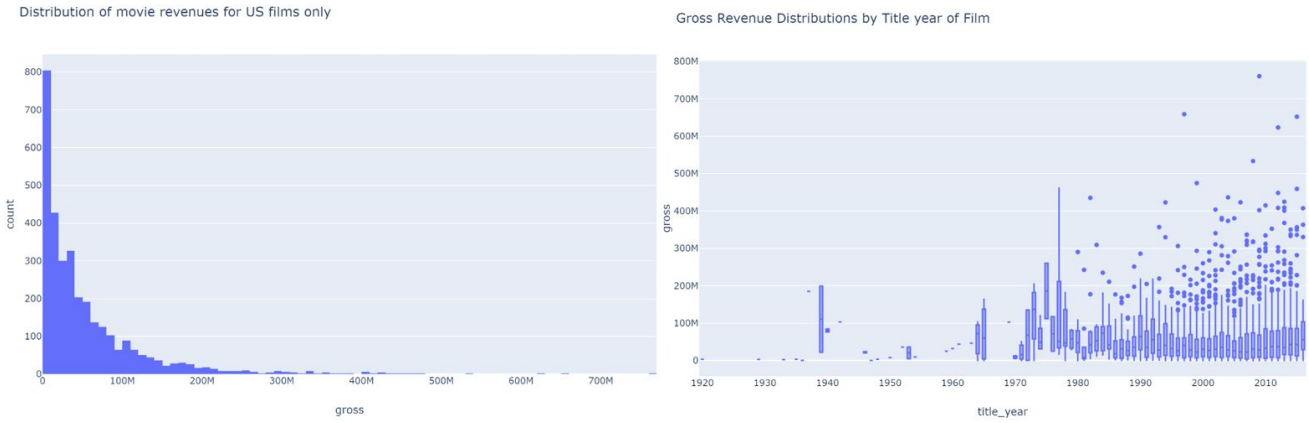

Figure 3: Overall distribution of film revenues is severely skewed right, distributions of film revenues are consistent from year to year.

2.3 Feature selection

Given the nature of the questions outlined above, our primary goal in selecting features from the data was to choose those that could best capture variance in gross revenue values, but did not relate to one another. This was achieved by a variety of means, the most straightforward of which was analyzing the relationship between some feature and gross revenue in the form of a scatter plot (**figure 4).** In addition to modeling the relations to visually choose features, we also performed a chi squared test using the null hypothesis that any particular quantitative feature had no relation to the target variable (gross revenue). Combining these two approaches, we settled on content_rating, imdb_score, budget, num_voted_users (how many people rate the movie on IMDb), movie_facebook_likes, cast_total_facebook_likes, and num_critic_reviews_for as possible features that related to gross revenue. We excluded features that would be highly correlated with one another. For example actor 1's Facebook likes are highly correlated with the cast's Facebook likes. Likewise we excluded from consideration features with little variance, such as movie duration (most films are around a standard length) and features with many unique and harder to quantify values (like the name of the director) as they did not provide any viable information for comparison between films. Since we also wanted to answer questions surrounding films' content rating in our project, we included content rating as a feature. In order to make analysis easier, we created a dummy variable for every unique content rating in the original column. Since we knew we planned on performing distance based algorithms like kmeans, we also performed a Z normalization on each column of the quantitative features.

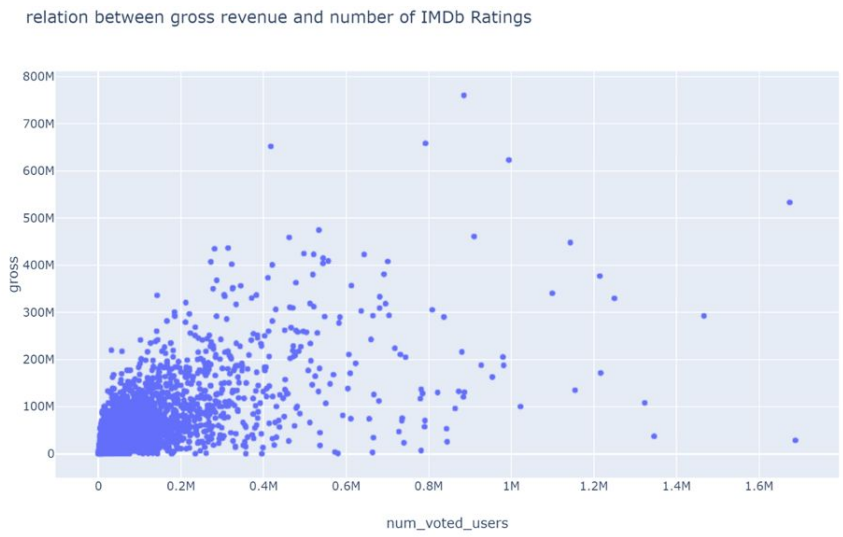

Figure 4: Using a scatter plot to analyze the relationship between two features

2.4 Supervised learning

We first tried to predict movies' gross revenue using a simple linear model. From our data explorations, we knew that our data were not completely linear, but we thought the accuracy of this model would serve as a good baseline for future iterations. We started off by training the linear model on all features initially identified. Somewhat unsurprisingly, this was not a very good way to predict revenue, resulting in a training R-squared score of 0.58 and a validation score of 0.53. Given the amount of features that went into this prediction, we knew that some of the features we were predicting on must have been insignificant, thus reducing the validation score. Using the results of the OLS regression output and the strength of the relationships between features we noted while performing exploratory data analysis, we successively dropped insignificant features and looked to achieve a lower BIC score each time. In this fashion, we achieved the highest values of R-squared across the training and validation sets when predicting gross revenue based on a film's IMDb score, budget, number of IMDb ratings, the sum of its cast's total Facebook likes, and the content rating of the film. The residual plot of this regression is shown in **figure 5**. In performing this regression, we noticed that our predictions were relying heavily on data generated from web interactions and social media (IMDb and Facebook for example). Thus, low accuracy might be due the incongruity of this data for older films and newer films that could benefit from the existence of these services during their release. To see if we'd be better able to predict the revenue of newer films, we did the exact same process

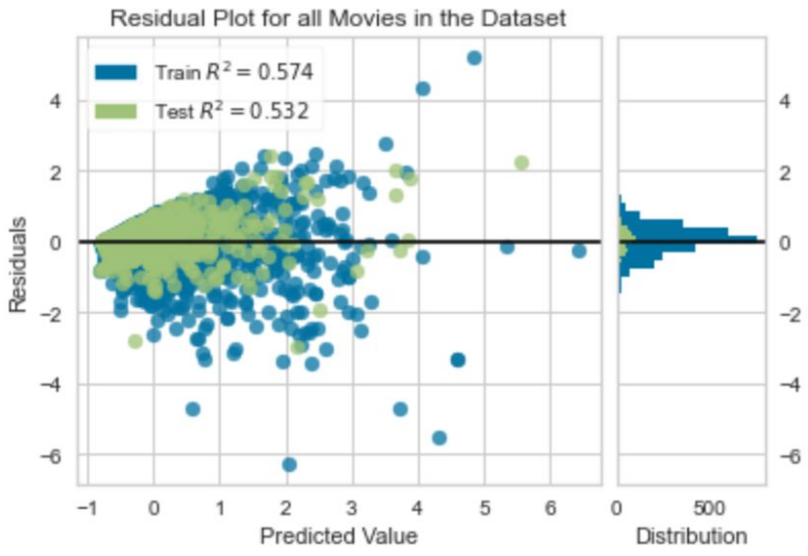

Figure 5: The residual plot associated with the most optimized linear regression for predicting movie gross revenue from other known information.

mentioned above, this time confining ourselves to films made in 2010 or later. What we found was an almost immediate 10% increase in model accuracy, even before optimization. After optimization we saw training and validation scores of R-Squared around 0.65 -- a noticeable improvement. What's more, in this regression, the OLS results showed IMDb score to be a more significant predictor of gross revenue than before, which affirms that perhaps these metrics are more accurate, and thus better for prediction, when dealing with newer films. However as noted in **figure 5** the residual plot for this data showed a relatively high degree of fanning, indicating the underlying fact that our data were not linear.

To address this problem, we tried to instead use random forests, which would hopefully be more versatile in predicting revenue and make fewer assumptions about the underlying data. Here we found that reducing the number of features based on which we were making predictions had no appreciable impact on the model's accuracy. Instead we optimized the model by trying to find the maximum tree depth that reduced overfitting the most and maximized correct predictions. We did so by plotting the error of the model (Training score - testing score) at various depths and selecting the best one (**figure 6**). This method indicated an optimal maximum depth of 8 branches for the random forest on all the films, and 7 for that on just newer films. In both cases however, the mean squared error of the predictions was about 0.23 (this is a normalized value for gross even though normalization is *not* specifically necessary for random forests). This shows an improvement from our linear models, whose mean squared errors ranged from 0.25 to 0.30, on the whole evidencing that random forests provided a better means to predict gross revenue than a simple linear model.

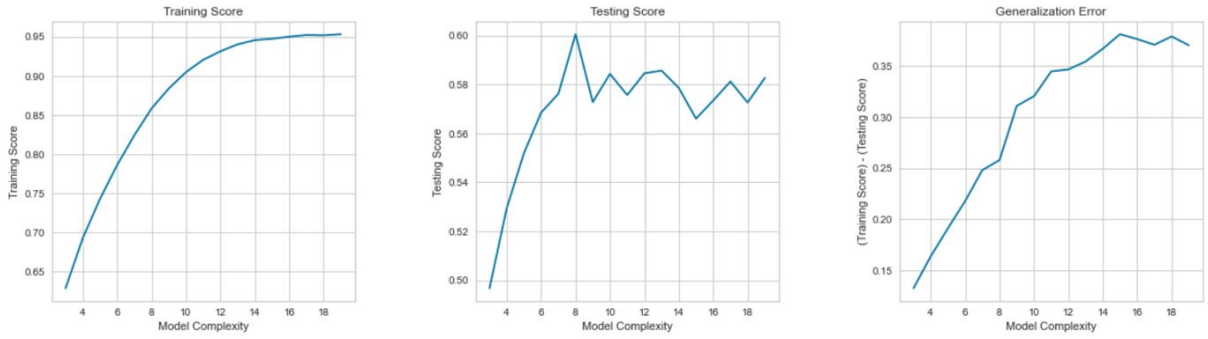

Figure 6: Visualizing the accuracy of Random Forest Regression to predict gross revenue at varying maximum tree depths.

The final aspect of supervised learning we implemented was a kNN algorithm to see if we could classify a movie's content rating based on the other information known about it. We felt kNN would be a good choice for this since our previous explorations showed that our data weren't really linearly separable, as some other classification algorithms assume. Here we again trained two seperate kNNs, one for the entire set of movies, and one for just those made after 2010. Two different approaches were used to predict the optimal number of neighbors. The first was similar to that outlined for Random Forests above, where the dataset was split into train and test sets, and the difference in accuracy scores between train and test sets observed over many values of k. We also did another assessment of model accuracy, this time using 10 fold cross validation to partition the data and then trying to minimize the classification error across different values of k (**figure 7**). When viewing the data in its entirety, the optimal number of neighbors was found to be 47, while for just the subset after 2010, 35 neighbors were ideal. At these levels of k, the model correctly predicted the content rating of new data 56% of the time for all films, and 59% of the time when restricted to new films alone. The associated confusion matrices are shown in **figure 8**. As shown in the matrices, the model was much more successful in classifying R and PG-13 films than G and PG films.

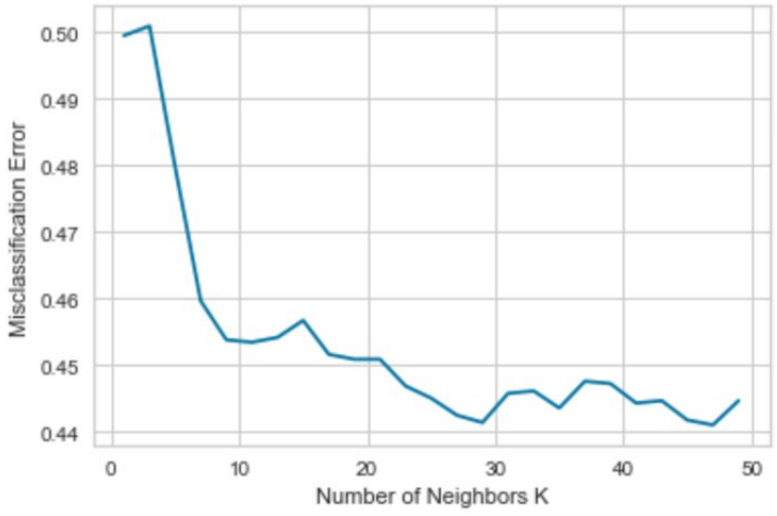

Figure 7: Minimizing misclassification error to find the optimal value of k in kNN.

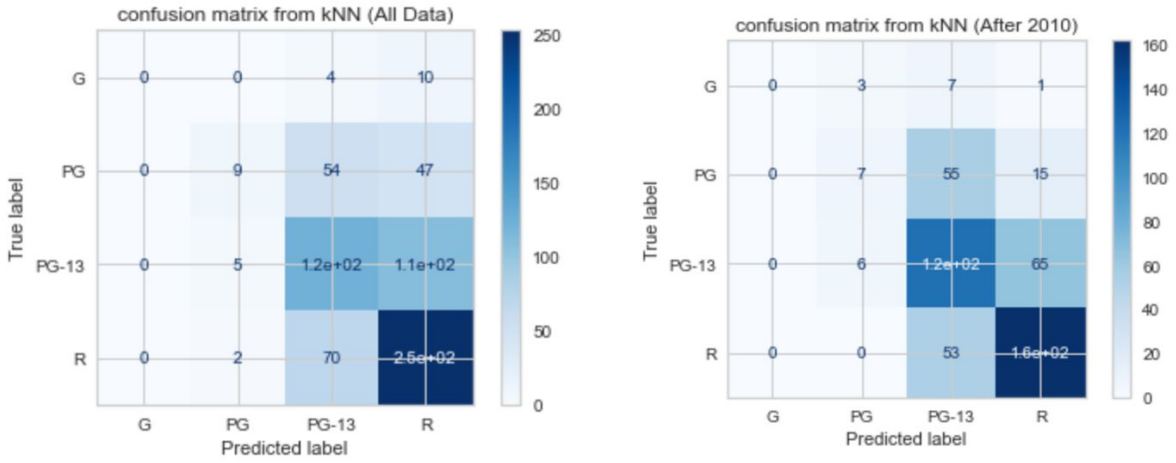

Figure 8: Confusion matrices for classifying movie content ratings. The number in matrix box *i,j* represents the number of films of type *i* classified as type *j*.

2.5 Unsupervised learning

We used the kmeans unsupervised learning algorithm both as a means of confirming our prior observations in the absence of ground truth labels, and to see if there was any logical significance to the way movies were grouped. This would potentially help distinguish between highly grossing films and the rest. To perform kmeans clustering, we'd first have to determine the number of clusters that best-suited our data. This was done using an elbow plot, and it was determined that we should form 3 clusters (**figure 9**). We then performed kmeans clustering on the data with three clusters. Given the added difficulty in deriving meaning from unlabeled data, and the fact that our clustering involved many dimensions, we decided to focus our analysis of the clusters around their distributions of revenue values, since that was our primary variable of interest. To do this, we extracted the movies that were placed into each cluster and their respective values for gross revenue, plotting the distributions in parallel boxplots (**figure 10**). What we found was that when looking at the entire set of films, and just those released after 2010, the same pattern emerged. One cluster of the three had a median gross revenue that was around 1.5 standard deviations above the dataset mean, while the other two clusters had approximately equal gross revenues, about 1 standard deviation below the mean.

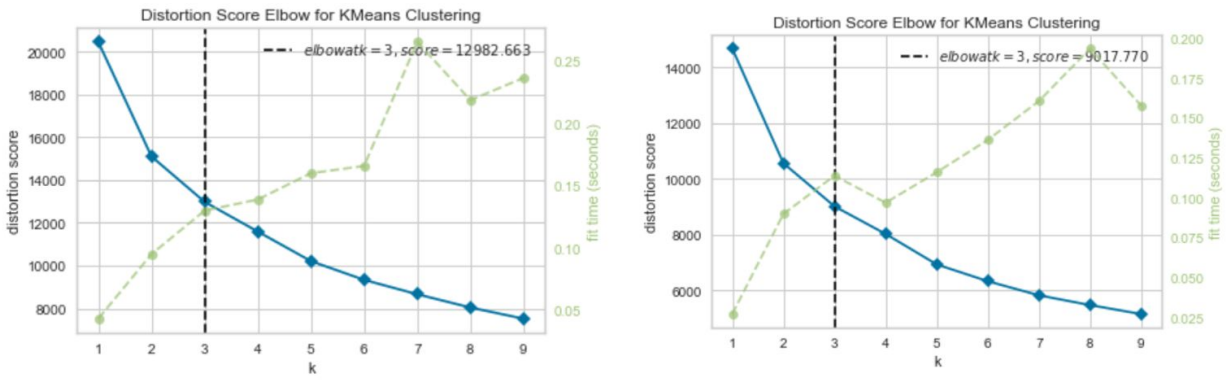

Figure 9: Elbow plots for kmeans clustering on all movie data (left) and movies made in 2010 or after (right). In both cases, the choice of three clusters is optimal.

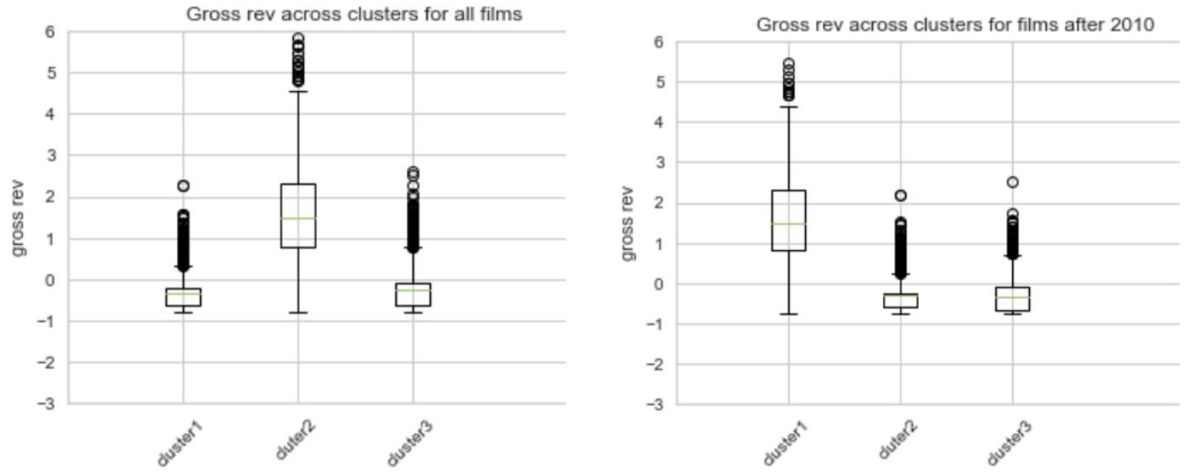

Figure 10: Distributions of gross revenue compared across clusters as determined by kmeans. Note: cluster labels are arbitrarily assigned each time clustering is performed, so the same cluster label between these two graphs could refer to two different subsets of films.

3. D3 Design decisions

Another part of our project, and of the learning process in our onboarding workshops in general, was learning how to display our data and findings interactively and intuitively using d3.js. For this purpose, we chose to make two interactive visualizations, static images of which are given below (**figures 11** and **12**). Each of these visualizations is related in some way to viewing or informing one of our central questions. Figure 11 shows a histogram with a variable number of bins. The histogram depicts the distribution of gross movie revenues in our data and changes as the user changes the number of bins. We thought this was a useful functionality because it allows users to see how skewed this data is when viewed at a high level of granularity (many bins) but how it can become more even when

taking a broader view. This is the same thing we noted in our earliest data explorations, where we saw that movie revenues as a whole were extremely skewed, but year-by-year they were more consistent. To ensure that the number of movies in each bin can be viewed visually as well as numerically, we added hover-over to each bin to display the number of films it contained. Figure 12 shows a boxplot, at first blank. When the slider is adjusted, it is overlain with points, each of which represents one individual movie and its associated revenue. When the slider is adjusted, the points spread out or contract so they can be either more easily differentiated from another, or so that the underlying plot can be seen more clearly. The dropdown menu allows a user to view the gross revenue distribution of movies of a different content rating and adjust the jitter of the overlaid points as mentioned previously. We created this visualization based on our initial observation that as films become geared to smaller audiences (G-PG-PG13-R) their average gross revenue diminishes. We felt that having each of the four revenue distributions on the same graph simultaneously would be too crowded to be visually appealing, and accordingly added the dropdown menu to allow the user to make selections. Since this meant that the scale of the plot would be changing, we added hover-over features that would display the dollar amount of a particular movie's revenue when the corresponding point was moused over.

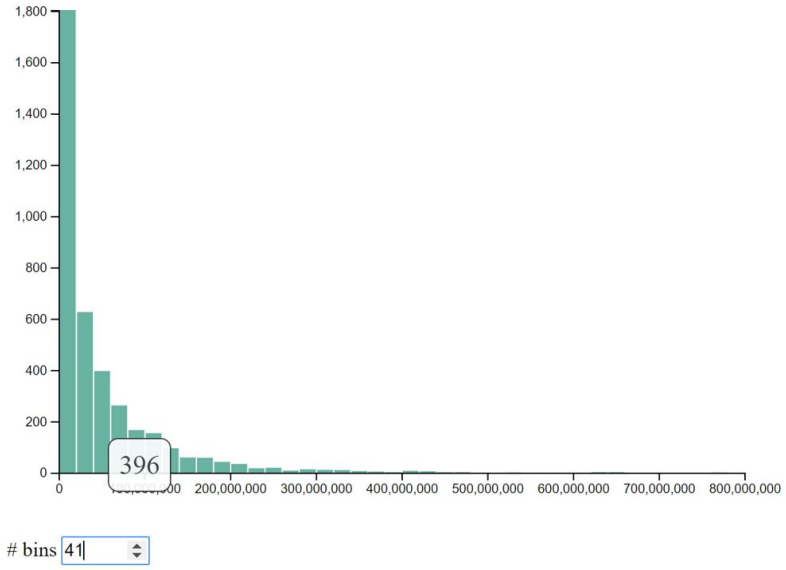

Figure 11: Static view of a d3 visualization for the distribution of gross revenue in our data. The number of bins can be incremented and incremented, and hovering over a bar shows the number of films (count) in that revenue range.

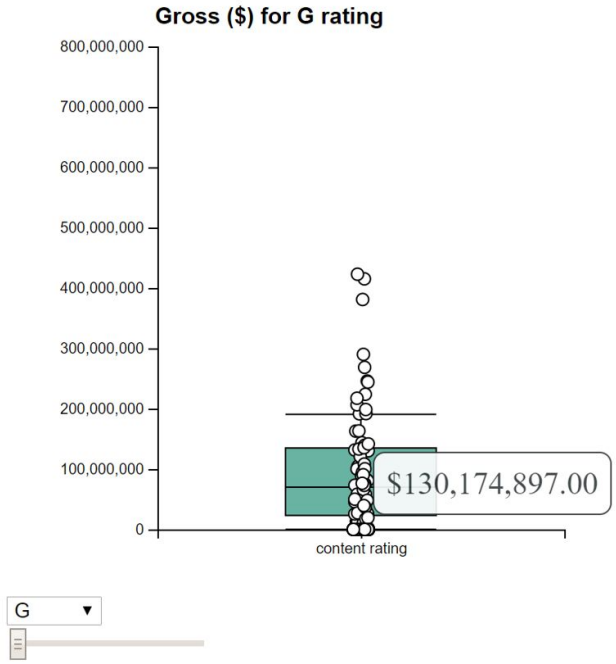

Figure 12: Static view of a d3 visualization for movie gross revenue Grouped by content rating. Hovering over a point displays the associated Revenue value. The content rating displayed and the jitter of the points can Be adjusted

4. Conclusions and future work

In our project we aimed to investigate two things: the relationship between a film's engagement characteristics and its revenue, and the predictability of a film's content rating (this was to see if we could accurately learn about the content of the movies given other features). We started our analysis by performing a simple multiple linear regression on the set of features we had determined to be most related to our target variable, gross revenue. When we noticed the most relevant features were those involving online and social media involvement (Facebook likes, IMDb ratings, etc), we thought it prudent to form a subset of the newer films so that these features would be more representative of how people currently respond to and interact with films. Doing so, we saw an almost 10 percent increase in the R-squared of the linear model. Even though the residual plots show that the data is best *not* linearly approximated (unless we were to transform the values of gross revenue somehow) it demonstrated that more accurate predictions can be

made about movie revenue if using films that have always existed in the era of social media and the online world. In trying to make a more flexible model to predict movie revenues, we optimized a random forest regression on the same subset of features as in the linear model. With the optimized random forest, we saw less overfitting than we did with linear models, and we saw an overall increase in accuracy (a decrease in mean squared error) from the linear models. Here too the regression was more accurate on the subset of newer films, but not to the same degree as with the linear model. One feature of these data we noticed is just how skewed the distribution of gross revenue values was. This is most likely why the more flexible random forest was a better predictor than linear regression, but neither were extremely successful. This skewed shape of the distribution is highlighted in our d3 visualization, and also in the unsupervised clustering analysis we did on the data. We used kmeans clustering to group our data into 3 clusters across each of the features we identified earlier. Even though clusters were based on all of these features, revenue between the same clusters showed a very similar pattern: one cluster contained films far above the dataset average gross revenue, the other two clusters contained films far below the dataset average gross revenue. This seems to indicate that movie revenues don't fall on a gradient so much as they are either 'average' or extremely high. Looking at the data on a micro scale, we can see many of the highest outliers seem to be franchise films, those that repeatedly have people coming back. For these films especially, that have such a large following, we'd *expect* their social media and internet fan engagement to be high in proportion to revenue generated. In general however, we were able to predict with some accuracy movie revenues based on features like IMDb score, the cast's Facebook likes, number of movie ratings, etc, and this accuracy increased when we confined our analysis to only newer films. This points to these metrics as being viable predictors of movie revenue, and perhaps other measures of overall performance, as they are a readily available, and easily interpretable way to gauge user response (as opposed to something like a review itself for example).

We also sought to investigate the link between the content of a movie and how people responded to it based on the simple metrics we had. Our approach to this question was to use kNN to classify a film as one of the four major content ratings.

We thought this could be an interesting application because of the link we saw between content rating (a measure of audience breath) and gross revenue. Ultimately, with the limited level of detail provided by our feature set, classifying content ratings was a more difficult problem than we anticipated. There was little difference in model performance between the new films subset and all of the data. In addition, as shown in the confusion matrices, it was hard to distinguish between PG-13 and R movies, and PG and G films. This result makes sense because of the similarity of each of these pairs to each other. Since most of the features we used serve to reflect how users engage with a film, the difficulty in classifying based on these features could mean that user engagement does not fall along lines of content rating.

It is also important to address some concerns and limitations associated with the type of analysis we performed. Most prominently is the risk of comparing currency values that are not obtained on equivalent scales due to inflation. This is a valid concern because our data do include movies from as early as 1910. To investigate this issue further, we read about how IMDb aggregates its currency information. IMDb itself states that it reports gross revenue figures as a cumulative total. Looking further, the general IMDb site obtains its gross revenue values from a service called “Box Office Mojo” that is associated with the premium, professional tier of IMDb. Box Office Mojo say on their page that when amassing info on movies, they do adjust for inflation over time when evaluating ticket prices. Although this allows us to be relatively sure that our analysis is not jeopardized by inflation, one potential solution could be to normalize all of the gross revenue values to a standard Z distribution for all those movies from a particular year. That way, a film’s normalized gross value reflects how well it competed against only those films released in the same year. Interestingly, when looking at Box Office Mojo, we also discovered additional metrics that are available to industry professionals, such as gross revenue within a movie’s first week of release. Although we don’t have access to such data, they could make for a compelling analysis because they are a better indicator of initial public response.

As we have mentioned already, another limitation, or perhaps caveat of this dataset is that by nature, films released many years before the advent of the internet (let alone social media) simply won't have the online followings that newer films do. We tried to account for this by segmenting the data used in our analyses, and we found that the more recent the film, the better the regression is able to predict on its features. It is important to note that we chose the year 2010 quite arbitrarily, and that a more statistically informed choice of this 'cutoff year' could produce even better results than those we observed. Even still, the problem of differing context will always be a concern in such an analysis. For example, it may be the case that modern films have a greater chance of having large Facebook and IMDb followings but conversely, movies released many years ago benefitted from the greater popularity of movie theaters at the time. This is obviously much different now when movies can be streamed, and attendance to theaters is down.

In the future, we'd like to incorporate even more interactive elements to let a user experience the results of the various statistical and quantitative methods we employed, but on the whole we thought this project was an interesting way to execute a data science workflow and gain insight into even one of the most pedestrian and well-known parts of life: movies.