Capstone 1 Consolidated Report

Problem Statement

My capstone project aims to build a tool capable of successfully predicting the final domestic box office gross of a film with high accuracy and precision.

Data Wrangling

To build the dataset required for a predictive model, I acquired data from multiple sources and combined them to create my movies dataset. Each source brought unique features that would assist in predicting the final box office gross. The base dataset was The Movies Dataset, available on Kaggle, combining the information of over 45,000 films in the MovieLens dataset with 26 million user reviews. This dataset had several important features such as genres, languages spoken, release dates, however, it featured little to no information on the box office performance of the films. Amongst the few films that did have box office figures, only worldwide gross numbers were provided. While international box office gross serves an important role in recovering a film's budget, comparing the box office performance of two films can be challenging. Not all films released in the same number of countries at the same time, the independent small budget films often do not play at all in countries where language can prove to be a barrier. Action-oriented blockbusters do play well amongst audiences with a limited grasp on the English language, there is minimal dialogue and an abundance of action set pieces that can be universally relished in such films. Amongst big event films, the release dates can often be delayed by months in different regions. Despite the recent growing focus on worldwide box office performance, primarily due to China's quickly expanding market, for older films, worldwide box office numbers are either unavailable or inaccurate due to figures from all markets not being thoroughly tracked. These reasons outline the choice to focus on the United States of America and Canada's box office numbers, referred to as the domestic box office gross from here on out, over worldwide numbers. Beyond that, the base dataset while providing user reviews had no consensus for critics' reaction to the films present in the dataset and thus the need for other sources became readily apparent.

The OMDB API was used as a source for data, returning all the data present for any given film in either XML or JSON, depending on the response type chosen by the user. The OMDB API included scores from Rotten Tomatoes, a website aggregating positive and negative reviews from professional critics and returning the percentage of positive reviews as a score. Metacritic scores, an average score out of 100 of a film's ratings received from critics, was also available for many

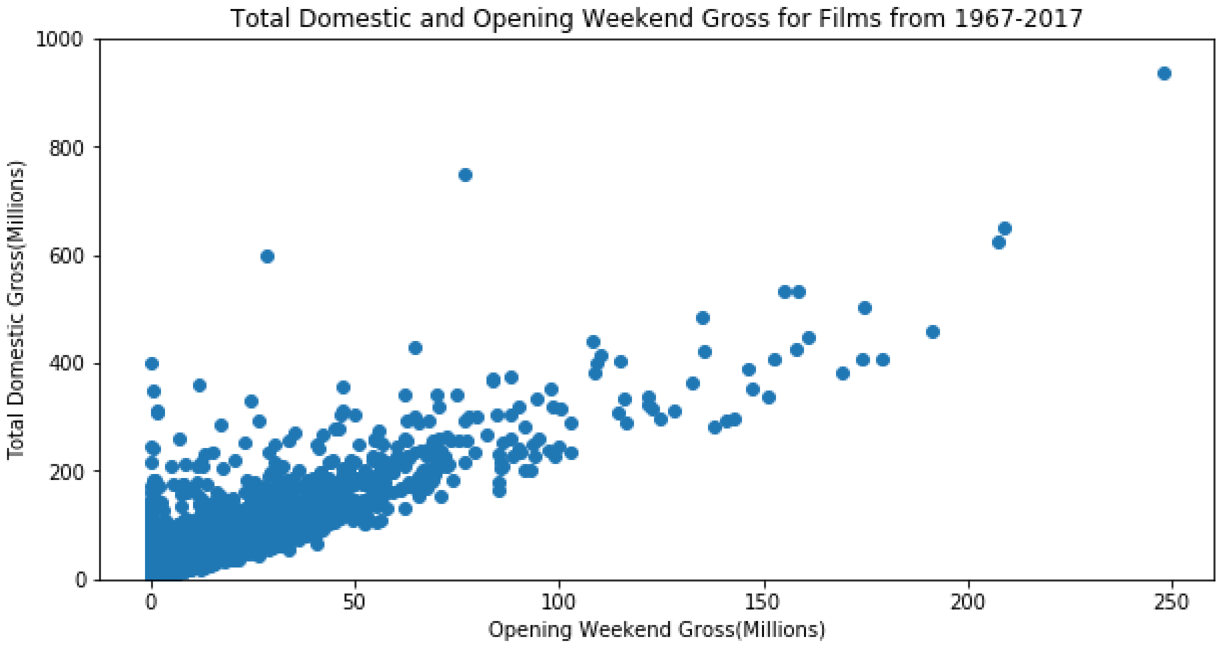
films. While the base dataset did provide audience ratings, OMDB API also provided the Internet Movie Database (IMDB) ratings, the average score given by users of the website out of 10, alongside the number of votes cast. These features would all prove vital in understanding the relationship between the audience and critics' reviews and the domestic box office performance of a film. Alongside the critical and audience reception, the OMDB API provided other purposeful features such as MPAA ratings, a label provided to indicate to theatres what the appropriate age is for the viewing of any given film. I used the IMDB ID, an alphanumeric primary key for IMDB's movies, available in both the base dataset and in OMDB API as a request option, loaded the OMDB data for all the films in the base dataset. While OMDB API had a domestic box office feature, gross for about only a tenth of the base dataset were available, the need for another source for box office figures was evident.

Box office data was sourced from Box Office Mojo, an industry leader in reporting box office figures. Box Office Mojo provided along with final domestic gross numbers, opening weekend gross, the gross collection between a film's first Friday to Sunday window and production budget, cost of the film barring marketing and promotional expenses. For the domestic gross, only the box office business during its initial release is considered. Although infrequent in recent years, prior to the existence of home video, popular films would often receive a limited re-release as besides being broadcast on television, it was the only way for audiences to see a movie after its theatrical run at the box office. Re-releases are being ignored for the same reasons international box office performance is not being considered, it gives certain films advantages that others do not receive. If we were to take the 460 million dollars Box Office Mojo states the original 1977 *Star Wars* grossed, it leads to miscalculations for our predictive model. *Star Wars* overtook 1975's *Jaws* as the highest-grossing film of all time which held the record then with 260 million dollars. Except during its original theatrical run, *Star Wars* made 307 million dollars, a further 15 million came in a re-release in 1982 and over 138 million came in 1997 during the special edition re-release of the original *Star Wars* trilogy. The ticket prices rose dramatically during the 20 years between initial and special edition releases, the national average price for a movie ticket in 1977 was $2.23, in 1997 it was $4.97. *Jaws* never got re-released and so using the combined numbers inflates the box office performance of *Star Wars* during its original theatrical window at the box office.

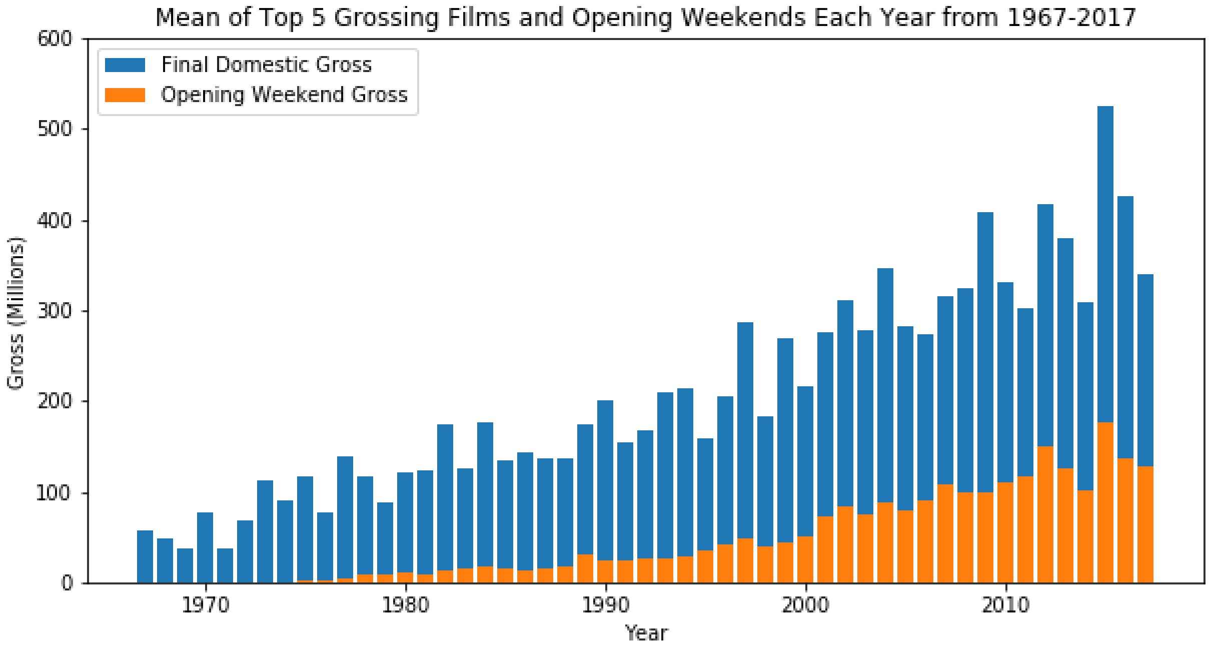
Due to Box Office Mojo and IMDB both being owned by Amazon, all the film's on the Box Office Mojo website had an IMDB ID attached as well. Using the IMDB ID provided from the base dataset, I requested all the information from the OMDB API for each film of 45,000 films in the original dataset, followed by scraping the Box Office Mojo page for each given film for the relevant box office data I required. I requested and stored the relevant additional information in a new pandas dataframe and then merged them using IMDB IDs for a new dataset exported to a CSV file. Once the movies dataset was created, I explored the data using a few built-in methods to understand the steps required to prepare the dataset for analysis. I saw that among the 45,000 films that were stored in the dataset over 12,000 had final domestic gross numbers and over 11,000 had opening weekend grosses, a feature that would prove crucial in the assistance of the predictive model. Opening weekend grosses are rarely substantially impacted by reviews, for word of mouth on a film to build, a large section of the audience needs to see the film first. Opening weekend is a greater reflection of marketing and promotional efforts than the critical reception a film garners. The financial figures studios invest for marketing and promotion are not released publicly like box office figures. The opening weekend is a tough number to predict even within the industry and large errors in predictions are commonplace. Having an opening weekend number as a base assists the model greatly in predicting final gross factoring in both critical and audience reception alongside performances from similar films and other metrics the dataset provides. IMDB ratings and votes are present for nearly all 45,000 films barring a 100, Rotten Tomatoes scores are available for slightly less than 20,000 and Metacritic scores for just under 12,000 films. Since nearly all the features of the dataframe were stored as string values, I converted the important metrics to their respectful types, monetary figures were transformed to float objects, release dates to DateTime objects and so forth.

Data Visualization

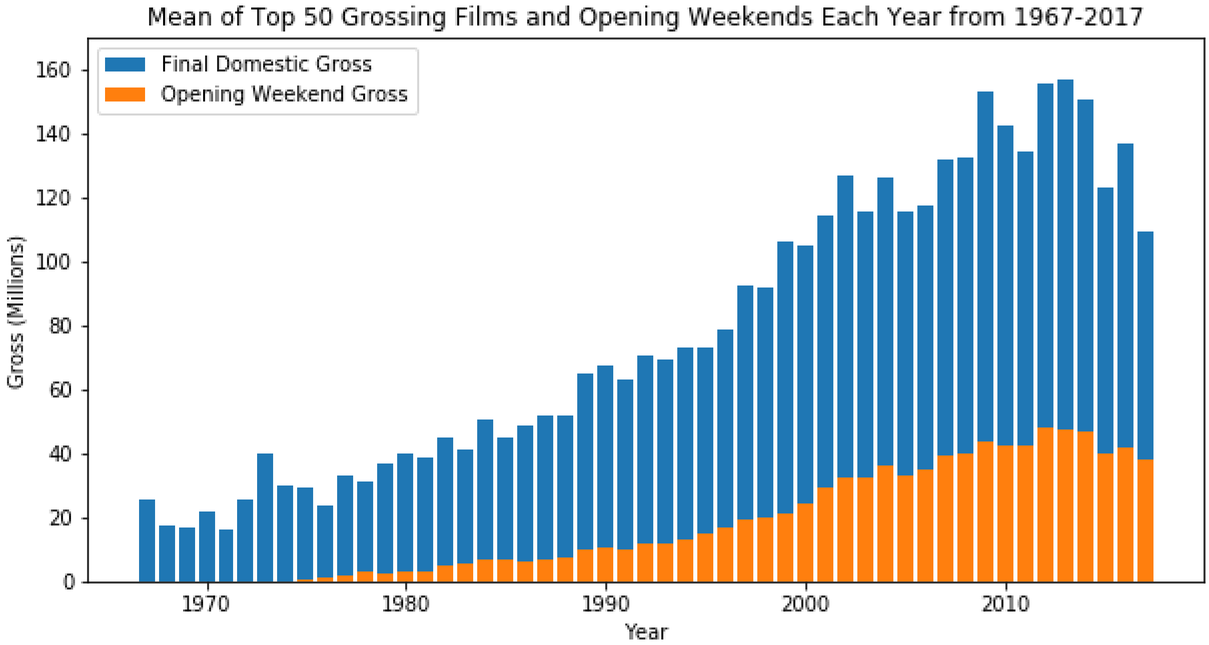
I created a subset of the original movies dataset, isolating only the films that had a domestic box office gross data This new subset contained 12,000 of the 45,000 films available in the original dataset, the movies without relevant box office data were discarded as they would add little consequence to our exploratory analysis. I exported the new box office dataset to a new CSV file, loading it into a pandas dataframe for both my data visualization and statistical analysis components. For the data visualization, I focused specifically on final domestic gross and opening weekend gross, particularly highlighting trends over time as the dataset contained films from the beginning of the 20th century to the penultimate month of 2017. Using matplotlib, I plotted the relationship of domestic gross and opening weekend gross for all films in the last forty years in my box office dataset below.



As can be observed from the graph above, there appears to be a relatively proportional relationship between the final domestic and opening weekend grosses for the films in the dataset from 1967 to 2017. Understanding this relationship will be paramount for the predictive model and hence looking at how this relationship has developed throughout the forty years being displayed became important. To avoid biased results yielded from only examining the top-grossing films of each respective year, I took the mean of the top five films of each year to account for the fluctuating nature of the yearly box office.

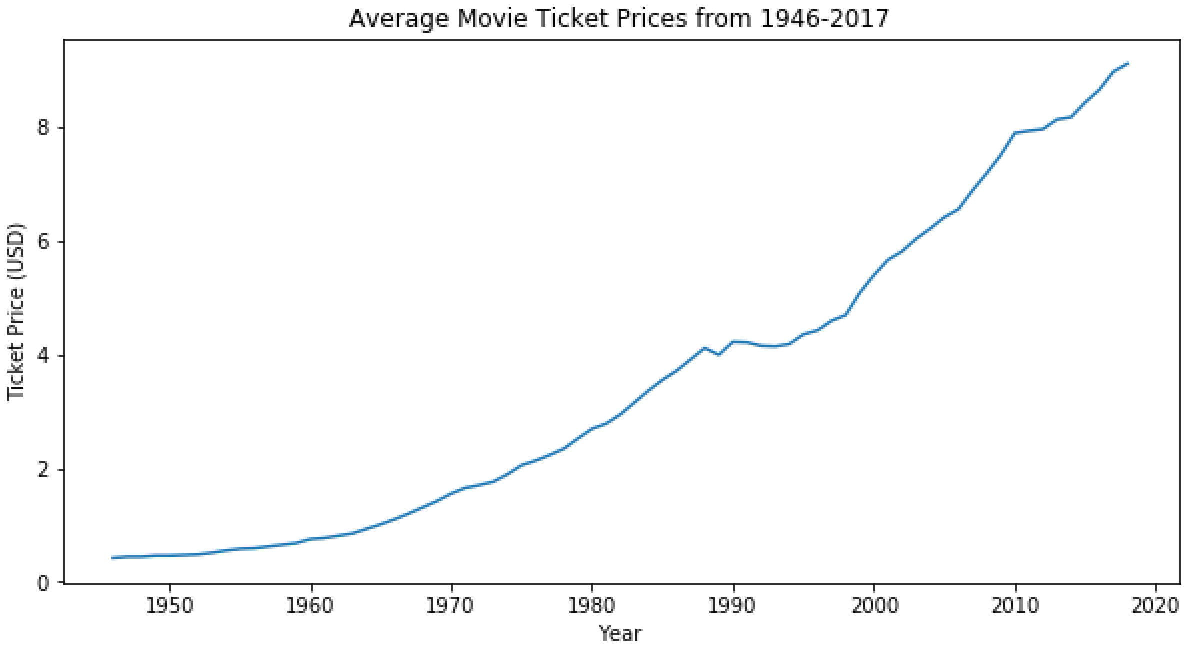


This graph represents the trend for films that were very successful in their respective years. The graph below takes the mean of the top fifty to get a better idea of the performance of the overall year instead of highlighting just the most financially successful films.

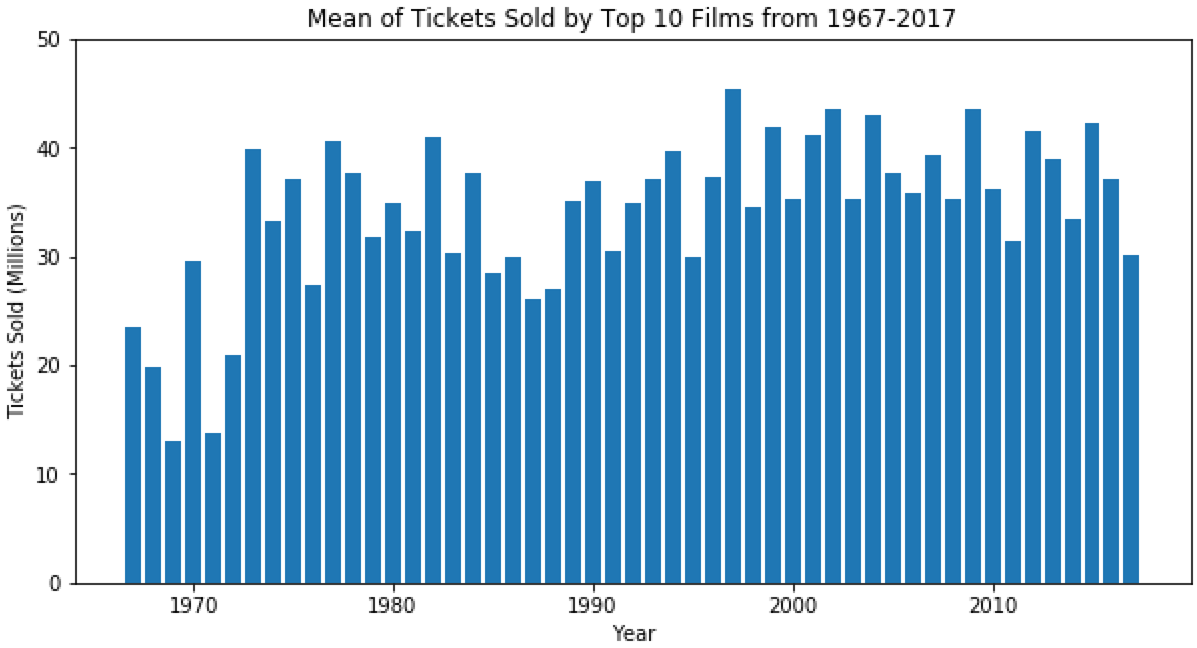


Immediately apparent is the rise in importance of the opening weekend grosses, in both graphs growing at a considerably faster rate than domestic gross over the last forty years. Also, notice that 2015 in the earlier graph, towers over the years surrounding it with more than 100 million dollars higher mean amongst the top five films. This was the year *Star Wars: The Force Awakens*, the highest-grossing film domestically of all-time with 936 million dollars, and then record holder of the highest opening of all-time with 248 million dollars was released. Also released that year was *Jurassic World*, which earlier in the year, broke the opening weekend record with 208 million on its way to a final domestic gross of 650 million dollars, which in any other year in the graph above barring 2009, would have been the highest-grossing film of the year. The extraordinary performance of these two films inflates the top five mean of 2015, observe the graph above displaying the mean of top fifty films of that year, omitting 2017, 2015 has the lowest mean in the last ten years.

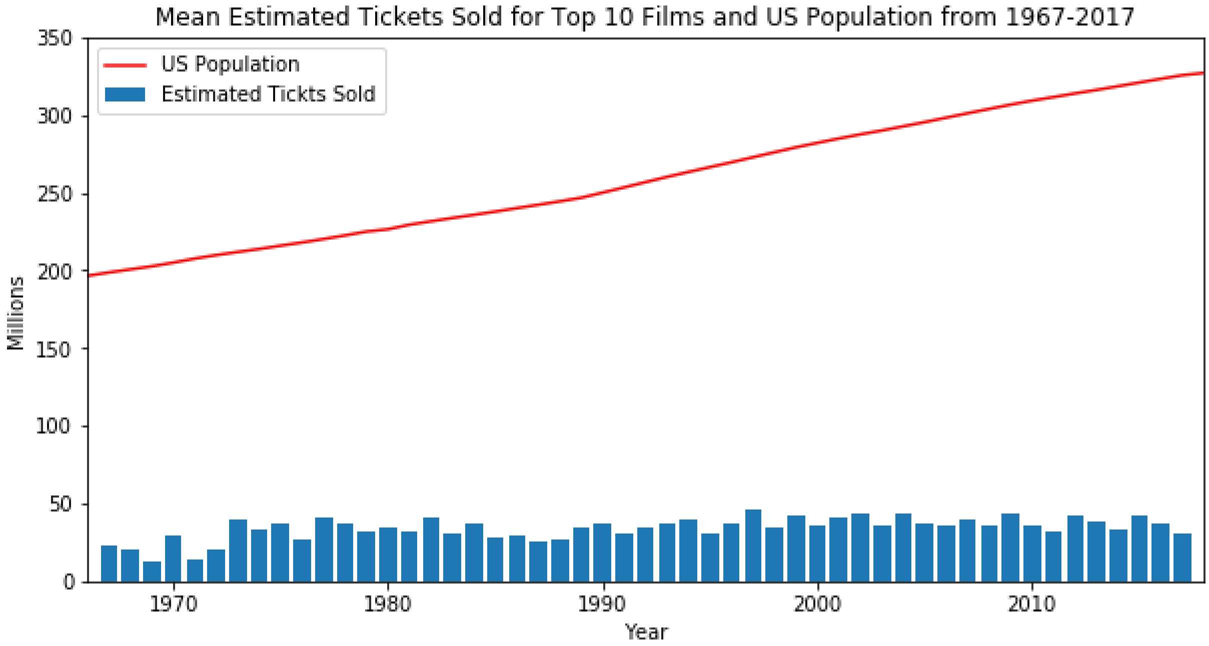
While we see positive growth in both opening weekend and final domestic grosses for films in the last forty years it is important to consider the rise in ticket prices over this period of time. When the original *Star Wars* released in 1977 the ticket price on average was little over 2 dollars compared to over eight and a half dollars for *Star Wars: The Force Awakens*. The graph below displays the rise of the national average ticket prices from 1946-2017.



The graph above shows a fairly consistent rate of increase in ticket prices to account for inflation, with relatively few years in which ticket prices did not rise from the previous year. Using these ticket prices, we can now estimate how many tickets were sold for all the films in the dataset. This is an estimate, ticket prices often vary in different regions and there is no data provided to compare a film's performance in different states or cities to better estimate accurately the number of tickets a film has sold. Ticket prices in urban areas are often far more expensive than rural areas. Newer films often release in different formats such as 3D or IMAX which enables theatres to charge a few dollars extra per ticket for most big blockbuster releases. The graph below is an estimate of the mean of tickets sold by the top ten films at the box office every year from 1946 to 2017.

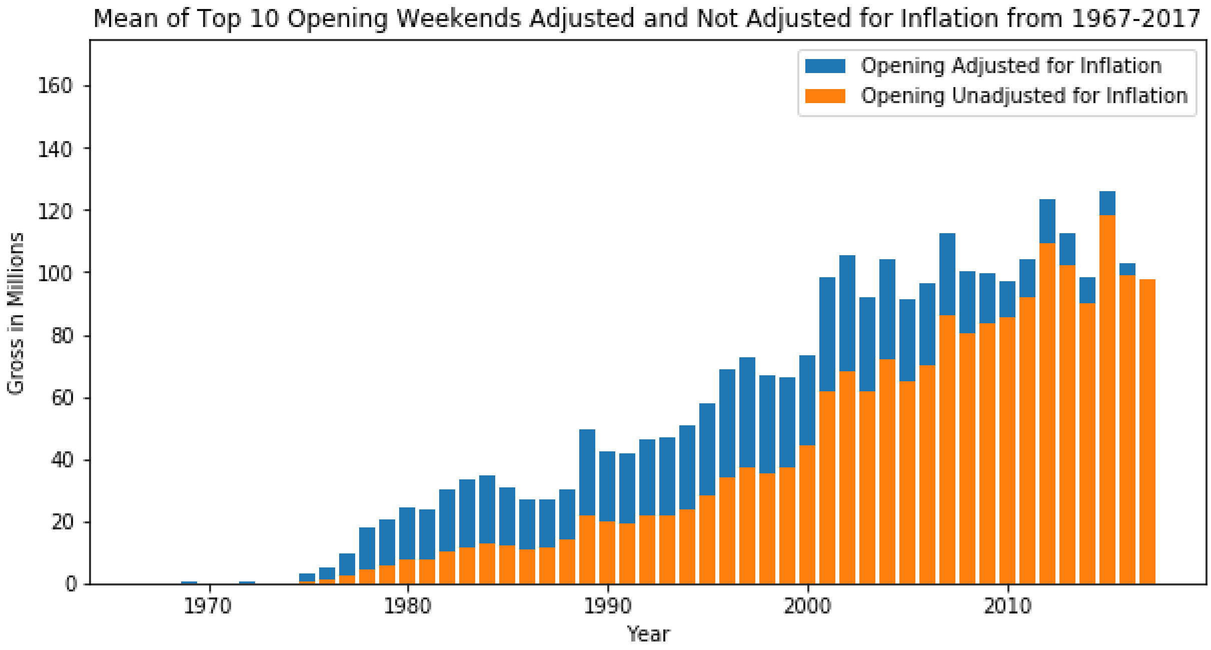


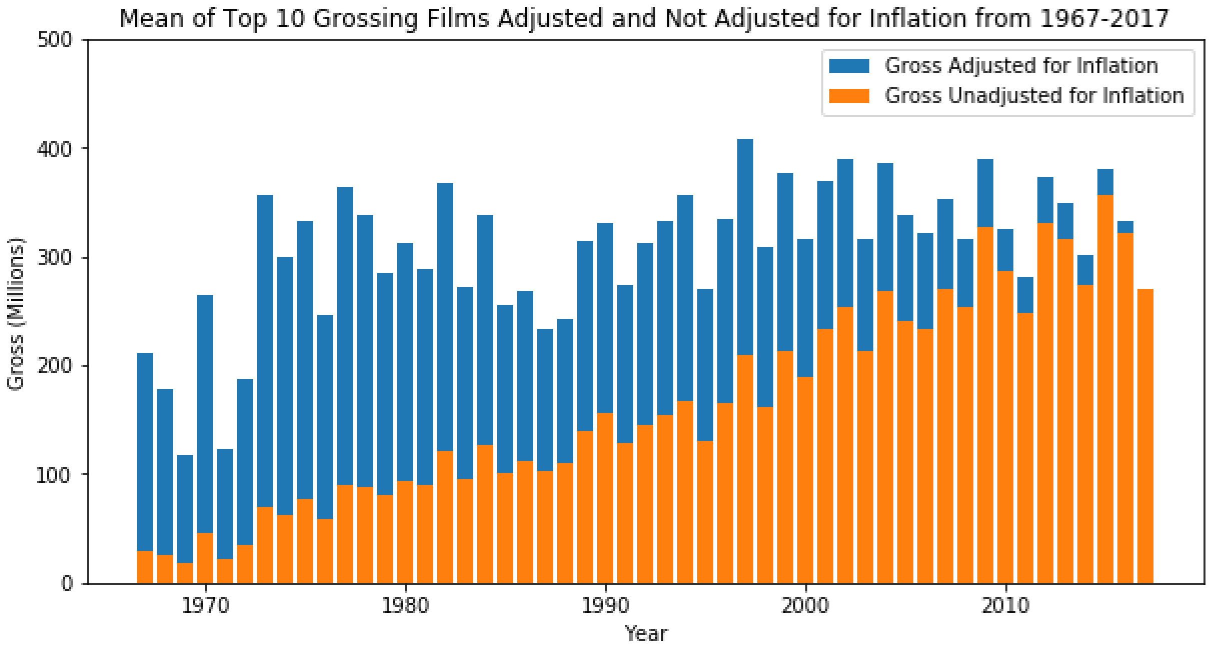
We can observe a relatively consistent number of tickets being sold in this period of forty years, a majority selling over 30 million and under 50 million tickets. There is a larger disparity between years early in the graph, partly due to box office figures for many films in that era not being available. While the graph above may suggest that box office performance has remained consistent over forty years. The graph below displays the growing population in the United States during the same forty years in the graph above.



Another point to note for the graph above is that only the population of the United States is shown while tickets sold in Canada are also included in the estimates which are responsible for approximately ten percent of tickets sold. Where the tickets sold remain consistent as the prior graph suggested the population in the United States consistently rises year after year. At the beginning of the graph, the population is under 200 million whereby the end is approximately 330 million. While 40 million tickets sold would account for a fifth of the United States population in 1967, in 2017 that same figure would only account for only an eighth of the total population. The number of tickets sold has not seen any significant decrease over the last four decades, the percentage of the population that goes to theatres to watch films have seen a steady decline.

Finally, the two graphs below display the opening weekend and domestic gross at the box office of the top ten films of their respective years both adjusted and not adjusted for inflation, inflation accounted for by multiplying the estimated tickets sold to the average ticket price in 2017.

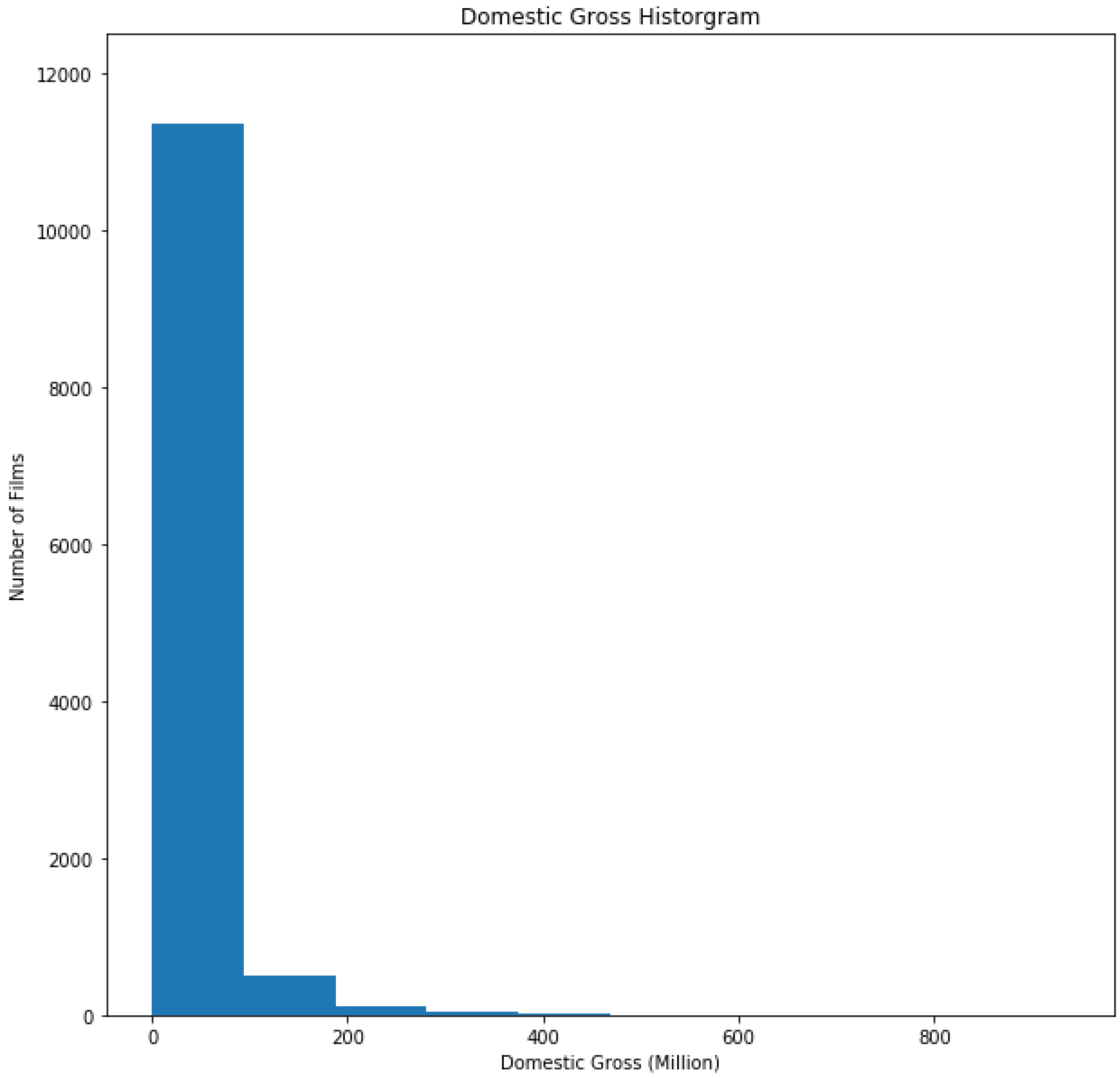




Both graphs validate our earlier claim, the opening weekend grosses have seen substantial growth over the four decades, while the growth of domestic gross is largely due to the inflation of ticket prices. There exists little opening weekend data for films of the 1970s and earlier, for those that do exist the numbers are significantly lower due to films in that era not opening nationwide, favouring a smaller release and expanding as word of mouth travelled. Most blockbuster films began opening nationwide from the 1980s onwards, even adjusting for inflation films in the 1980s rarely opened above 40 million dollars, whereas today the top ten opening weekends of the year open regularly above 100 million dollars.

Statistical Analysis

For the statistical analysis on my dataset, I decided to examine the relationship between both critical and audience reception of films to their box office performance. The graph below displays the histogram for the domestic box office gross from my dataset.

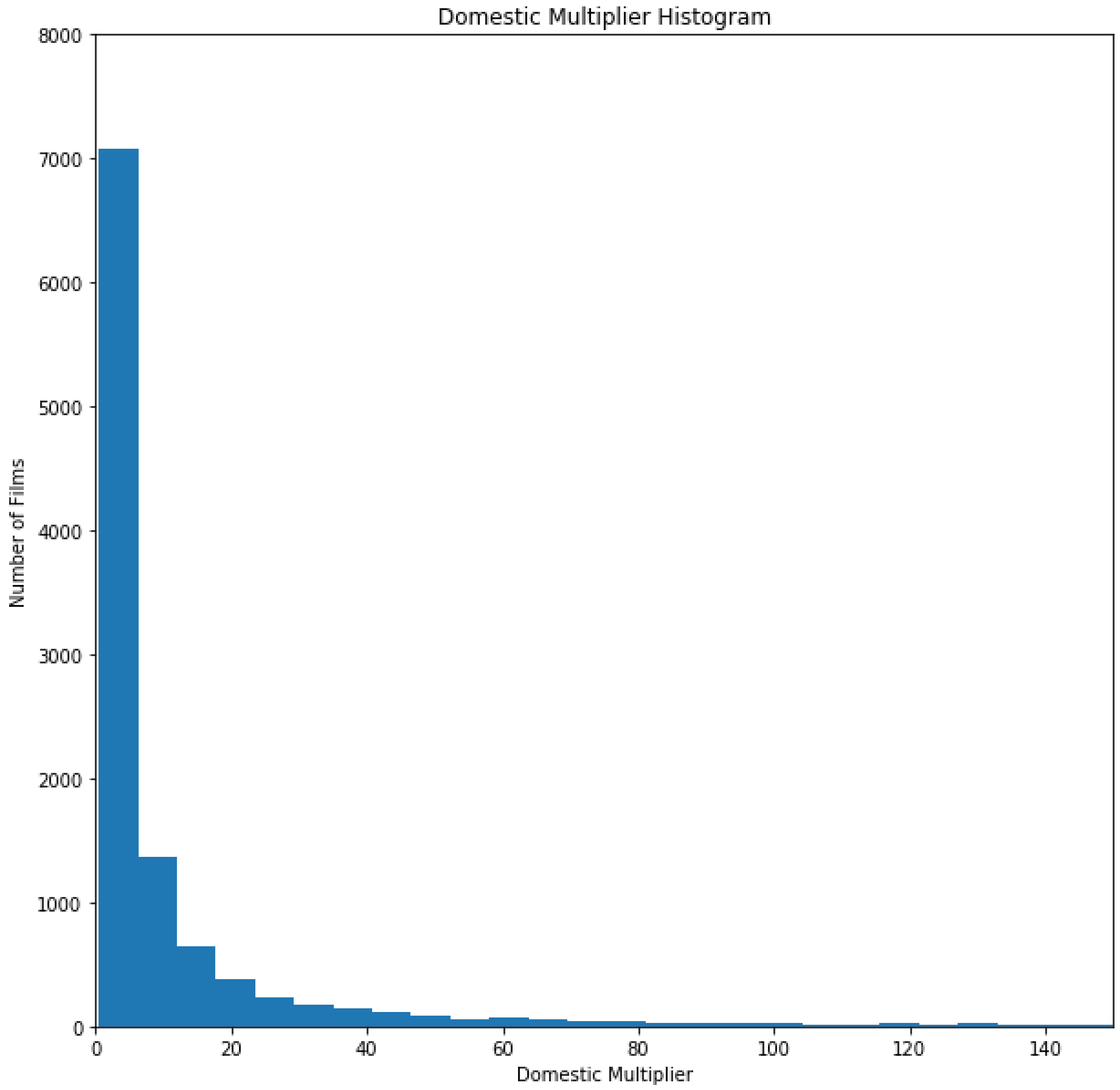


The majority of the domestic gross was under 100 million dollars, with the number of films grossing above 400 million dollars virtually invisible in the histogram above. Alongside the histogram I used the pandas' built-in methods, describe, on the box office dataframe which gave valuable insights. The mean of the domestic gross column in my dataframe 21.5 million and the standard deviation 47.2 million, confirming the high variance of the data. The metrics used to measure critical reception would be the Rotten Tomatoes score, for audience reception the IMDB ratings would be used.

The first null hypothesis was, the box office gross of a film and its Rotten Tomatoes score were independent of each other. To separate positive critical reception from negative reception, two subsets of the box office dataframe, one with films garnering a Rotten Tomatoes score of 80 or above, the other with films garnering a score of 40 or below were created. The boundary thresholds were each 20 from the mean Rotten Tomatoes Score, 59 for all films in the dataframe. Each subset contained around 3,500 films. Films with mixed reception were ignored, not assisting in our primary goal of discovering whether positive or negative critical reception affected the box office gross of a film. Using inferential and frequentist statistical techniques and over 10,000 replicates of the positive and negative reception dataframes, the p-value for the initial t-test was 0.29. The final p-value calculated using the bootstrapping approach with 10,000 replicates and the mean difference of the positively and negatively reviewed films was 0.0849. The p-value was larger than the 0.05 needed to be statistically significant and reject the null hypothesis. This outcome was in line with my original predictions, despite the critical reception of a film there exists an inherent built-in audience for certain films, primarily big-budget blockbusters and that audience is not swayed by critics reviews.

In 2016, the winner of Best Picture at the Academy Awards and holding a 98 percent on Rotten Tomatoes, Barry Jenkins's *Moonlight* opened to only 400,000 dollars in its opening weekend, it must be noted it only played in a handful of theatres in the metropolitan cities in the United States before going nationwide to 650 theatres and grossing 1.5 million during that weekend eventually making 28 million dollars by the end of its run at the box office. Unlike most blockbusters, films that studios believe will garner critical acclaim open in a small number of theatres before expanding hoping that the word of mouth would benefit its box office when it goes nationwide alongside box office boosts near the awards season if it musters several nominations and is seen as a serious contender for the major categories This is another reason films a studio believes are strong awards contender release near the end of the year to keep word of mouth strong and create a buzz prior to award shows that take place in the initial months of the year. The approach for event films such as *Batman V Superman: Dawn of Justice*, also released in 2016 is the opposite. Reviews for such films have little impact on its opening and so the strategy is to get as many people in theatres as early as possible, so a simultaneous nationwide and often worldwide release is favoured. Despite a dismal Rotten Tomatoes score of 28 percent, being panned by a majority of critics, the *Batman V Superman: Dawn Of Justice* opened to 166 million dollars opening weekend. This opening weekend alone accounts for approximately 6 times the final gross of Moonlight, the film eventually finished its theatrical run with 330 million dollars. More than half of its final domestic gross came in its initial three days even for films with such massive opening weekends they normally account for anywhere between 30-40 percent of its final gross. It is estimated the average household visits their local cinemas only twice a year and there simply is not enough room for the critically acclaimed independent films to attract the same audience as those lining up for big-budget blockbuster films. Validating the null hypothesis box office gross and critical reception are independent of each other.

Alongside final domestic gross at the box office, another important metric I wanted to use to measure box office performance was the domestic box office multiplier, a number dividing the domestic gross of a film by its opening weekend gross, a metric often referred to as the legs of film. The stark difference in the box office multiplier, with *Moonlight* having a multiplier over 10 and *Batman V Superman: Dawn of Justice* under 2 shows that critical reception can be impactful to box office performance. The opening weekend of a film is often a reflection of the marketing and promotional efforts of the studios and not the response from critics. Blockbuster films always will be marketed more aggressively than independent films, however, oftentimes films that have been panned by critics and audiences alike do not sustain the same momentum as a well-received film would at the box office and crash, leading to a relatively small multiplier. A large domestic multiplier is a sign that the film has been well received by audiences. The graph below is the histogram of the multipliers of the films in the box office dataframe.



Similar to the domestic gross histogram a majority of the films are near the origin of the x-axis, however, unlike the domestic gross histogram, the data is spread out throughout the x-axis. The mean multiplier on the dataset was 22.75 and the standard deviation was 118, an even larger variance that can be observed by the histogram as well. The reason for such high variance is as mentioned earlier the change in release pattern for films in recent decades. The original *Star Wars* released in only 43 theatres in its first weekend in May 1977. During August of that year, it expanded to over 1000 theatres. With an opening weekend of 1.5 million, finishing its original run in theatres with over 307 million dollars resulting in a multiplier just under 200. In contrast, *Star Wars: The Force Awakens* opened in over 4,000 theatres to a record 248 million opening weekend, finishing with a domestic total of 936 million. Its multiplier of 3.77, while the original film's multiplier is 50 times as large, *Star Wars: The Force Awakens* still has the highest domestic multiplier for any film to open over 200 million dollars. This highlights the change in multiplier in recent decades, due to the growing opening weekends we observed earlier in the exploratory analysis.

The new null hypothesis was that domestic multiplier and critical reception were independent variables. Repeating the steps of inferential and frequentist statistics taken earlier with domestic gross the initial t-test p-value was 3.85 e-35, substantially smaller than our earlier test. Using the same steps as before our final p-value was 0.0, a statistically significant score lower than 0.05, therefore we can reject the null hypothesis. Unlike with domestic gross, we can conclude that the domestic multiplier is dependent on critical reception and not independent as our second null hypothesis presumed.

I repeated the steps taken once more substituting critical reception for audience ratings to examine whether there would be any change with its relationship to domestic gross and multipliers. For the audience reception subsets of the box office dataframe, films garnering a rating of 7.5 or over made up the positive reception dataframe, films rated at 5.5 or under made up the negative reception dataframe. Unlike Rotten Tomatoes scores, which were out of 100, these ratings were out of 10. The boundaries only deviated 1 point from the mean, 6.5 for IMDB ratings, as opposed to earlier due to the IMDB ratings having significantly less variance than Rotten Tomatoes scores. Each dataframe had over 3,000 films which kept consistency with the critical reception dataframe used for statistical analysis.

The null hypothesis involving audience reception and box office multiplier could with a p-value of 0.0 be rejected similar to the result with critical reception. However, the null hypothesis involving audience receptions and domestic gross also resulted in a p-value of 0.0. Where the critical reception the null hypothesis was validated, in this case, audience ratings and domestic gross were not independent of each other. The explanation for why this occurred is yet unclear but it supports the theory that audience reception in the context of box office performance is more crucial than critics' reception. Through our statistical analysis, we can conclude that domestic gross and critical reception are independent of each other, whereas domestic gross and audience reception are not. The box office multiplier of a film is dependent on both the critical and audience reception.

Modelling Preparations

After performing statistical analysis on my dataset, I wanted to prepare it for training with regression models. For my data visualization and statistical analysis, I had kept a few features that were not numerical for reference but for this exercise, I dropped the non-numerical columns and created a new column for the month of each film's release. This column was added as box office performance is impacted by different events throughout the calendar year. The initial months leading up to summer are in the film industry included as part of the summer box office season. The first week of May and to the last week of August is considered as the summer season for box office collections. Prior to July, most students are in school, box office grosses during the weekends are stronger than weekdays. Following the summer break, the performances during weekends are slightly lower and supported by far stronger weekdays. During the holiday season, opening weekends are far lower than the summer due to the winter break for students and families taking vacations, weekdays during this period perform the strongest during the entire year. This leads to holiday season releases to enjoy larger domestic multipliers and is not always comparable to the multiplier for films released during the rest of the year.

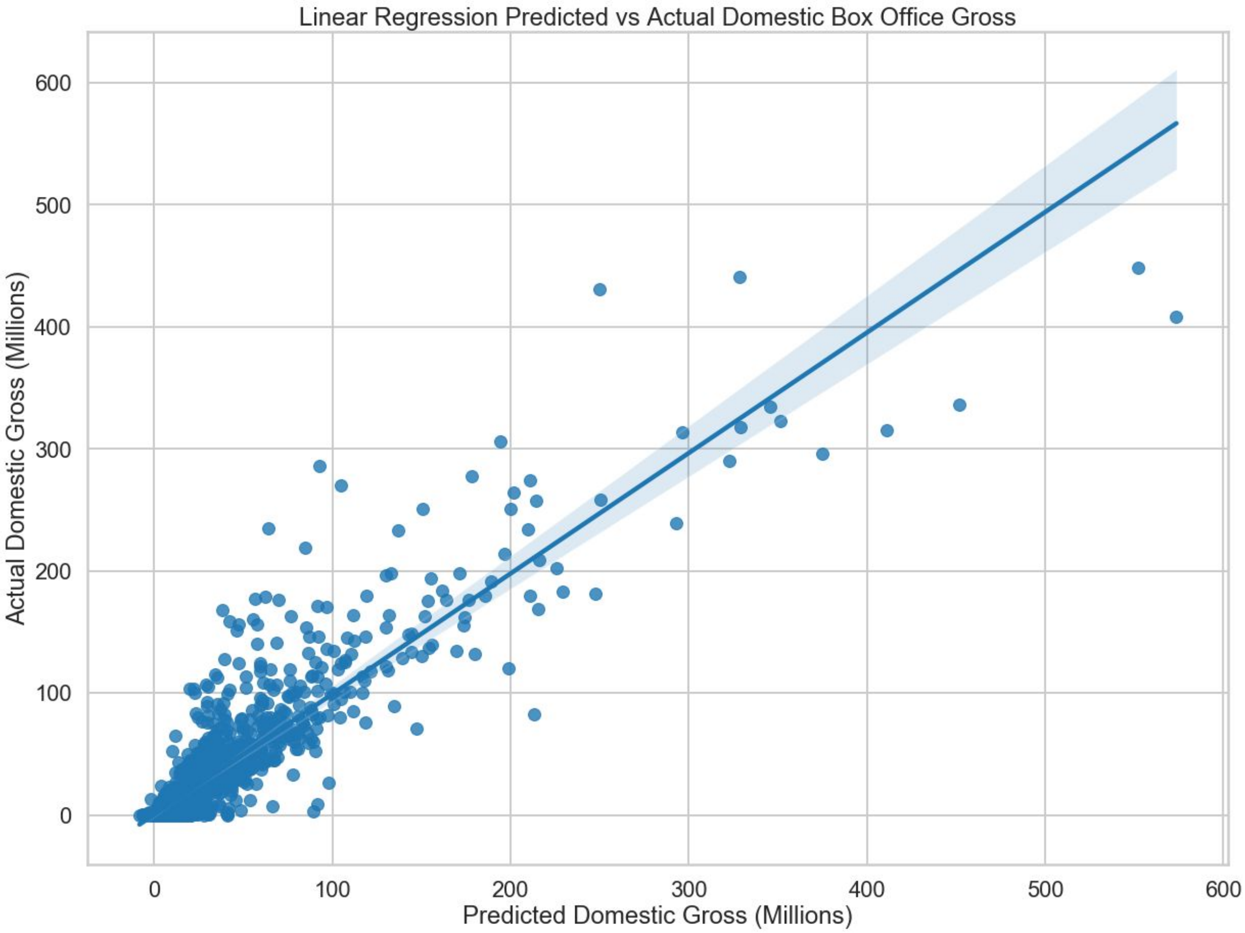
There has been a steady shift in recent years, large event blockbusters which released only during summer have started seeing more holiday releases. *Star Wars: The Force Awakens* was the first *Star Wars* film to not be released in May. When *Star Wars: The Force Awakens* opened *Jurassic World* earlier in the year broke the opening weekend gross record with 208 million dollars, no film at the time had opened to over 85 million dollars in December. *Star Wars: The Force Awakens* not only beat *Jurassic World*'s opening weekend with 248 million dollars but smashed the December opening in a single day, earning over 119 million on its opening day. Only three films since have opened to over 100 million in December, all three being *Star Wars* films. Therefore, the month of release becomes an important factor in analyzing box office performance and so it was added.

While most of the numerical columns had close to the 12,000 non-null entries, one column had significantly little data compared to the others. The budget column only featured a little under 4,3000 non-null values and since the next step in prepping my data involved filling non-null rows that would mean nearly two-thirds of the data in the column would not be real data and so I dropped the column. Once I dropped the unnecessary columns I used the fillna method choosing to replace null values with the mean of its respective column. Then I split my dataframe with the y variable being the domestic box office gross column values and the X variable including all other features that would be used to train the model.

Modelling and Results

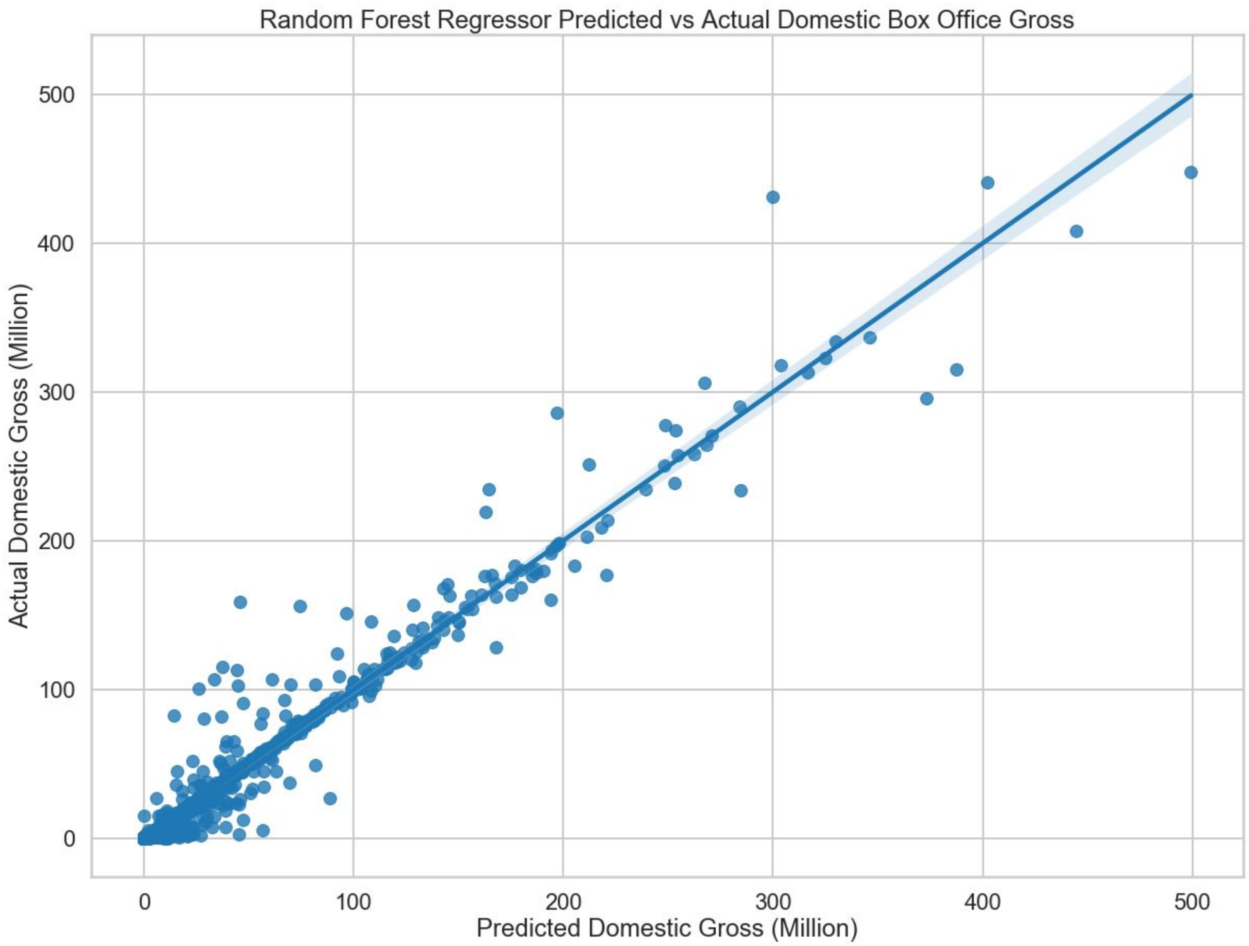
Once the preparations were complete, I first used an Ordinary Least Squares(OLS) regression model and observed the performance through the summary method of the OLS before trying more complex models and tuning the parameters. The R-squared value, the residual values squared for this model was 0.86. The residual is the difference between the actual and predicted values, squaring this number gives the same weight to positive and negative residuals. These values indicate a strong relationship and the purpose of the next models are to get an r-squared value as close as possible to 1.

I split the data into a training set and a test set using the train_test_split from scikit-learn's model_selection and set the size of the test set to 20 percent of the original dataset with over 2,400 films. This would be used to test our model's predictive capabilities on unseen data. The rest of the 80 percent would be used to train each of the regression models that follow. The first of which is a linear regression model with normalization set to true, assisting the model to deal with multiple columns with high variance in the data, as observed earlier from the statistical analysis. Once the model was fit to the training data, it was used to predict the y-variable in the test data, the final box office gross of each film. The plot below compares the results of the predicted values and actual values. To assist in making both of the plot's axes easier to read the predicted and actual values have both been divided by a million for all the graphs that follow.



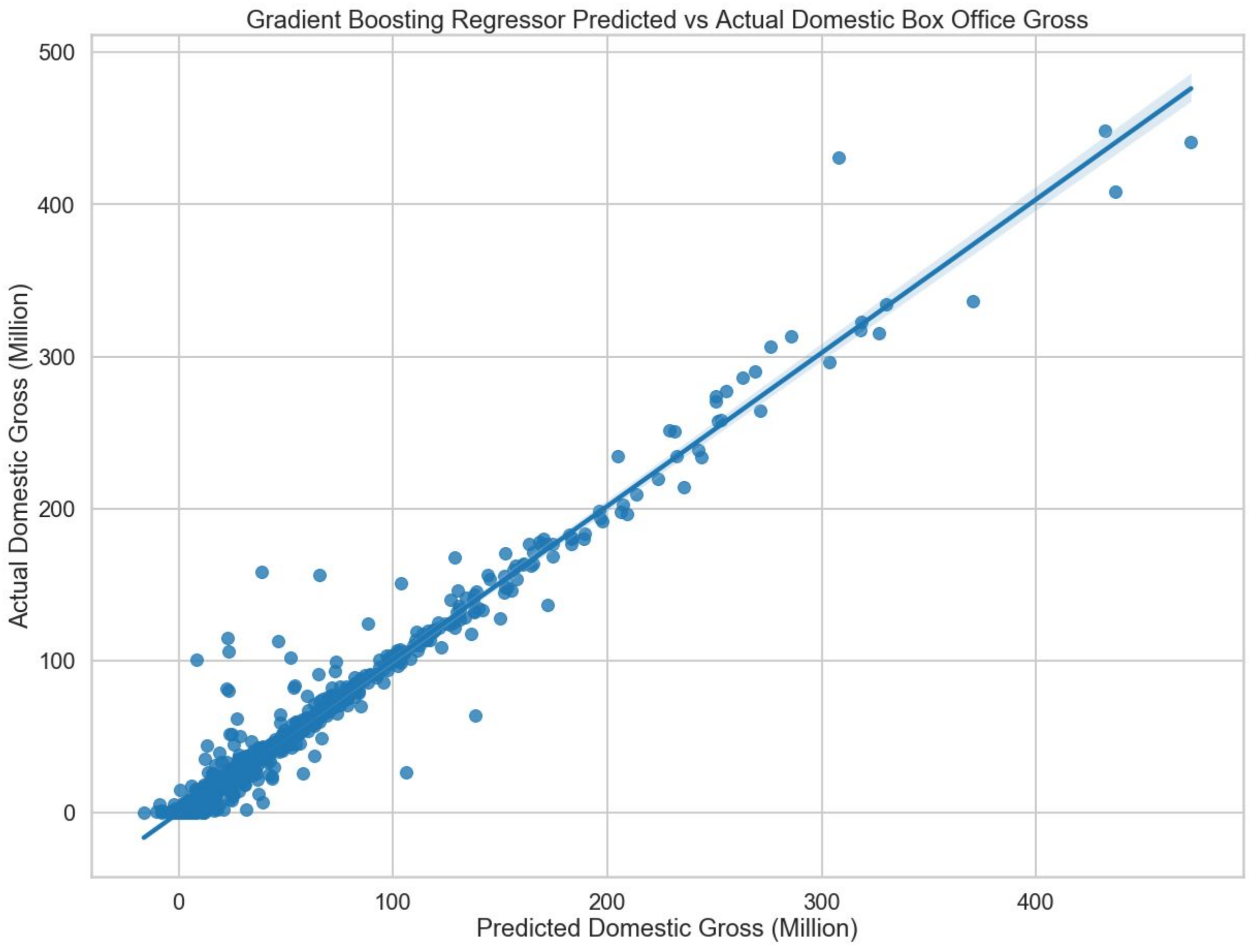
As observed from the graph above, the comparison does display a linear relationship but there do exist several outliers. The reduction of these outliers for the predicted and actual grosses on the graph in proximity to the regression line displayed indicates a better model. The model returned a similar r-squared value on both the training and test data, 0.8355 and 0.8217 respectively. These values being in close vicinity to each other suggests that the model is not overfitting. The mean y-value of the test dataset was 21.08 million, close to the 21.51 million dollars mean gross for the entire box office dataframe. The mean absolute error (MAE), a measure of the magnitude of errors in predictions, similar to residuals, taking the absolute difference between predicted and actual values, this model yielded an MAE of 8.89 million dollars. The root mean squared error (RMSE), a quadratic scoring rule taking the square root of the squared difference in predicted and actual values, resulted in an RMSE of 18.99 million for this model. Fivefold cross-validation was used to measure the performance of the model from different points in the training dataset, returning an average r-squared score of 0.8213. While the high r-squared values show the model is performing well to capture the explained variance and that the model is avoiding overfitting, high MAE and RMSE numbers suggest that the model is still not very accurate in its predictions and prone to large errors. Alongside the linear regression model, similar models using ridge, lasso, and elastic net regression were used and yield results indistinguishable from the linear regression model, they were not included in the final jupyter notebook visualizations.

The next regression algorithm I used was a random forest regressor using a 1000 n_estimators, the number of trees used in the forest. The random forest regressor was trained using the same training data used in the linear regression models. A new prediction for the y-variable in the test dataset was created employing the predict function of the random forest regressor. The graph below displays a similar comparison between predicted and actual values for the random regressor as shown earlier for the linear regression model.



The random forest regressor yields better results than linear regression in predicting domestic box office gross. A majority of the data points in this graph are substantially closer to the target regression line than the previous graph with the linear regression model. The r-squared values on the training and test data respectively are 0.993 and 0.9655, again validating the model is not overfitting. This confirms the initial observation from the graph that the random forest regression model considerably outperforms the earlier model. With the same mean for test data, 21.08 million dollars, the mean absolute error (MAE) is 2.01 million, a substantial improvement over the 8.34 million MAE for the linear regression model. Additionally, the root mean squared error (RMSE) is down from 18.99 million to 8.34 million for the current model. Fivefold cross-validation returned a score of 0.9285 compared to the prior model's 0.8213, all corroborating the current model appreciably outperforms the last.

The final regression model tested on the given training and test data was a gradient boosting regressor. Similar to the earlier model a n_estimator of 1000 was selected. The graph below presents the performance of the gradient boosting regressor comparing the actual and predicted values.

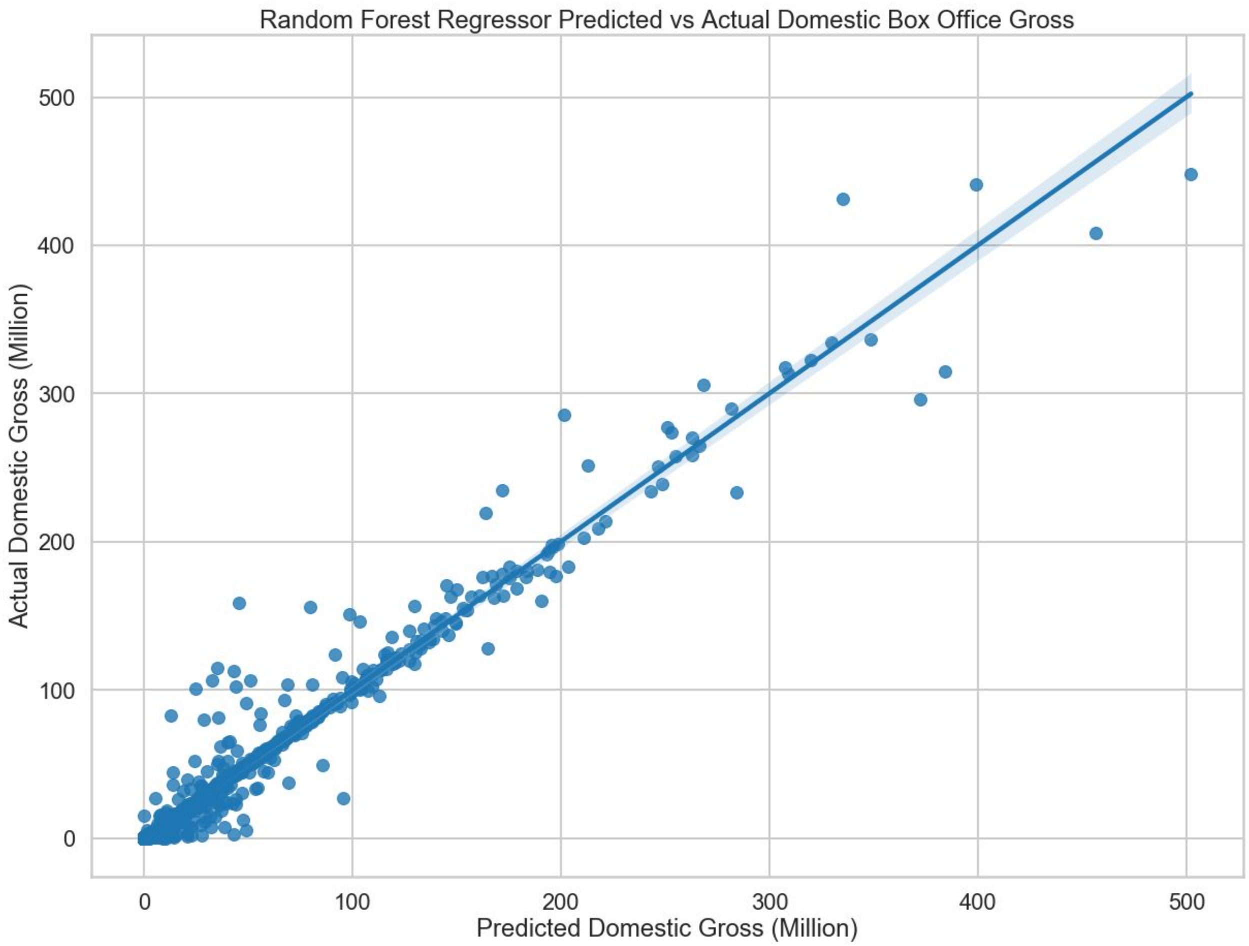


Visually the performance looks very comparable to the random forest regressor's graph. The data points particularly beyond 200 million dollars are better fitted, only one data point is distant from the ideal regression line. The data point is close to 450 million while the model predicts a gross closer to 300 million. It is the only such case, where visually we can see a prediction that is in stark contrast to the actual value beyond 200 million on the graph. The r-squared score for the training and test data is 0.9972 and 0.972. The model has avoided overfitting with the r-square values for both training and test data being relatively close. The r-squared values are slightly better than the random forest regressor. The MAE was higher than with the random forest model 2.39 million compared to the 2.01 million, but the RMSE performance was lower, 7.58 million compared to the 8.34 million. A lower RMSE indicates the model is less prone to large errors as they are additionally weighted for RMSE. The random forest and gradient boosting regressors performed well enough that both model's parameters were tuned to generate the best results possible to aid in determining the parameters for our final model.

Hyperparameter Tuning

For hyperparameter tuning, both randomized search and grid search cross-validation was performed on the random forest and gradient boosting regressor models. Starting with the random forest regressor, I used the get_params method to look at all the parameters available for the random forest regressor and determined meaningful parameters to tune. I created a random grid dictionary storing the parameters and the various values to assess. Due to the randomized search requiring a reduced compute time, only fitting 100 candidates with the cross-validation set of 3, regardless of the number of parameter values I tested far more values for randomized search than grid search. Once I had gotten the best parameters through the randomized search, I adjusted each parameter until finding the finest possible performance.

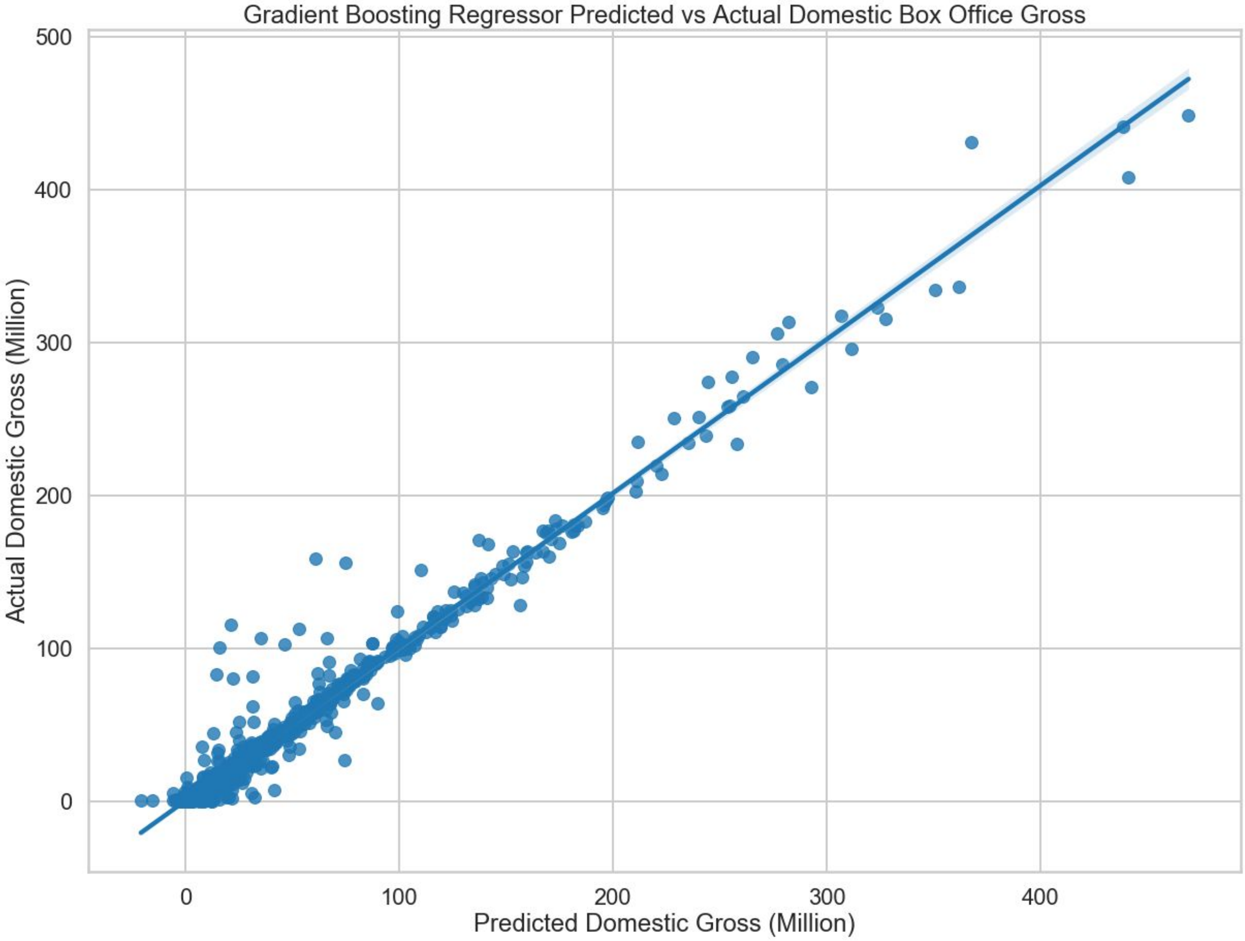
It was hard to distinguish any real improvement, the test data's r-squared value is 0.9677, improved by 0.0022 from the regular random forest generator. The MAE is 1.96 million compared to 2.02 million in the base model. The RMSE saw the most substantial improvement from 8.34 million in the original model down to 8.07 million with randomized search cross-validation. The fivefold cross-validation score was 0.9288 is marginally higher than the earlier 0.9285. Now that I understood the ideal parameters, I could test far fewer values for the grid search and then fine-tune the parameters to get the optimal performance from a random forest regressor on my data. After creating a new param_grid dictionary with 1050 candidates testing fitting 3 folds each, I ran the grid search cross-validation to find the best parameters. The performance of the new model using the tuned hyperparameters provided from the grid search cross-validation can be observed in the graph below.



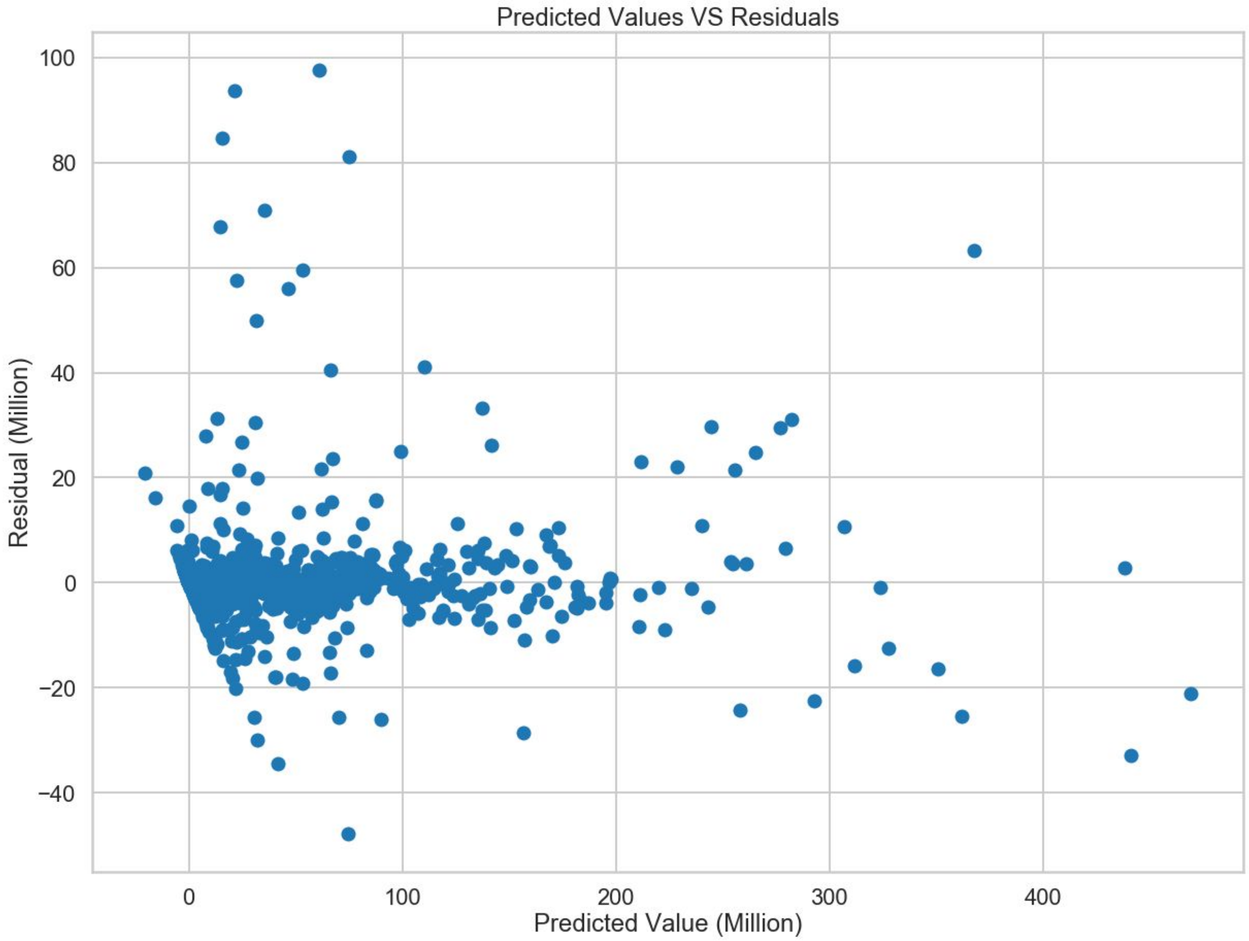
The performance is nearly indistinguishable to that of the parameters used for randomized search. The only significant improvement in this model compared to the last is the MAE, further reduced to 1.96 million where all the other metrics measured were nearly identical. While the random forest regressor has seen minor gains after hyperparameter tuning its performance largely is still below that of the base gradient boosting regressor model. After tuning the parameters using both randomized search and grid search cross-validation, the gradient boosting regressor would be the model of choice.

I repeated the steps taken for random forest regressor's hyperparameter tuning for the gradient boosting regressor, observing all the parameters could be tuned with the get_params method. I began with performing the randomized search cross-validation with largely the same random grid dictionary, removing the bootstrap parameter, which is not present in gradient boosting regressors and adding the learning rate. Once I ran the randomized search, I got the ideal parameters that were returned and made slight changes to all parameters searching for the ideal performance.

The r-squared score increased from 0.972 in the base model to 0.9765. The MAE has decreased to 1.91 million compared to 2.39 million earlier and is now even lower than the best random forest model. The RMSE also showed a healthy drop from 7.58 million in the base model to 6.89 million. The fivefold cross-validation score for the model was 0.9422 the best we have seen so far. I could curate the parameters I wanted to tune for the grid search cross-validation using the parameters returned by the randomized search. Using 3 folds of 1260 candidates my grid search totalled 3780 fits and returned the best parameters. The graph below shows the predicted values versus the actual values for my data.



This final model performs visually quite similar to the earlier model, a modest improvement can be seen in the test data r-squared score with 0.9786 compared to 0.9765. While there is a higher MAE value with 2.10 million, it still outperforms the base models MAE of 2.39 million. The RMSE dropped even further to 6.58 million from the 6.89 million RMSE of the previous model and 7.58 million of the base model. The fivefold cross-validation score returned was 0.9378, marginally lower than the last model. This model's lowest RMSE signifies the least amount of large errors, for predicting high variance data such as box office gross this final model has the best performance. Compared to the first linear regression model, where MAE was 8.89 million and RMSE was 18.99 million with the average gross for the test data being 21.08 million this model is a monumental improvement. The final graph below displays the predicted values compared to its residuals.



As observed from the graph above, the majority of large residuals are for films predicted to gross under 100 million dollars. Beyond 100 million, only a single data point can be seen, either positive or negative, larger than 40 million in the top right corner. Comparatively, there are 4 data points in the top left showing a residual of over 80 million for films predicted to gross under 100 million dollars. The predictive model often underestimates with significantly more positive residuals than negative; a single data point in the whole graph is lower than negative 40. This underestimation may be caused due to the domestic gross for all films being unadjusted for inflation. As observed earlier in the data visualization, the opening weekend gross numbers have seen a prodigious increase over the last few decades, while overall domestic gross adjusted for inflation remained largely stagnant.

Conclusion

The predictive model is highly accurate in its prediction of films with large grosses, most analysts and studios give greater importance to the box office performance of their big-budget blockbuster films, as opposed to the smaller budget films due much of the revenue being recovered through other channels namely selling streaming rights and award recognition. Therefore, the predictive model performs at its best for the values the clients of the model would be looking for, specifically, films predicted to at the box office have large opening weekends and domestic gross. For data that is as volatile alongside such high degrees of variance like box office gross, I am satisfied with the performance of the model and its predictive capabilities. In the future, there is still much that can be done with my dataset. While the gradient boosting regressor provides quite serviceable predictive modelling, the use of deep learning techniques such as neural networks to predict box office gross could yield greater results. The other avenue left for exploration is prediction without the use of the opening weekend gross. This was noted earlier as challenging, particularly without the cost of the promotional efforts of the studios releasing their films. Yet a model that can predict box office gross with some level of accuracy would be highly beneficial to movie studios as critical response and opening day audience reviews can be available a few days prior to the release. A model that would not require opening weekend data could predict the box office run of a film the day it opens rather than the Monday after the film's release. While the journey of collecting the data from various sources and splitting it for training and tests had resulted in an effective predictive model, there exists still, a great deal of improvement that can be made in both the results produced and shortening the time taken to yield meaningful predictions.