Capstone 1 Milestone Report

Problem Statement

This 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. While this dataset had several important features such as genres, languages spoken, release dates it featured little to no information on the box office figures of the films. Amongst the few films that did have box office figures, it was worldwide gross numbers which I wanted to avoid. International box office serves an important role in recovering a film's budget, comparing the box office performance of two films can be challenging. Not all films release in the same number of countries, the independent small budget films often do not play in countries where language can prove to be a barrier. Action-oriented blockbusters, however, do play well amongst audiences with a limited grasp on the English language as there exists minimum dialogue in such films. Even 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 film's worldwide box office numbers, they are often 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 apparent.

The OMDB API was used as a source for data, responding with 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 a percentage of positive reviews. Metacritic scores, an average score out of 100 of a film's ratings received from critics, were also available for many films. While the base dataset did provide audience ratings, OMDB API also provided the Internet

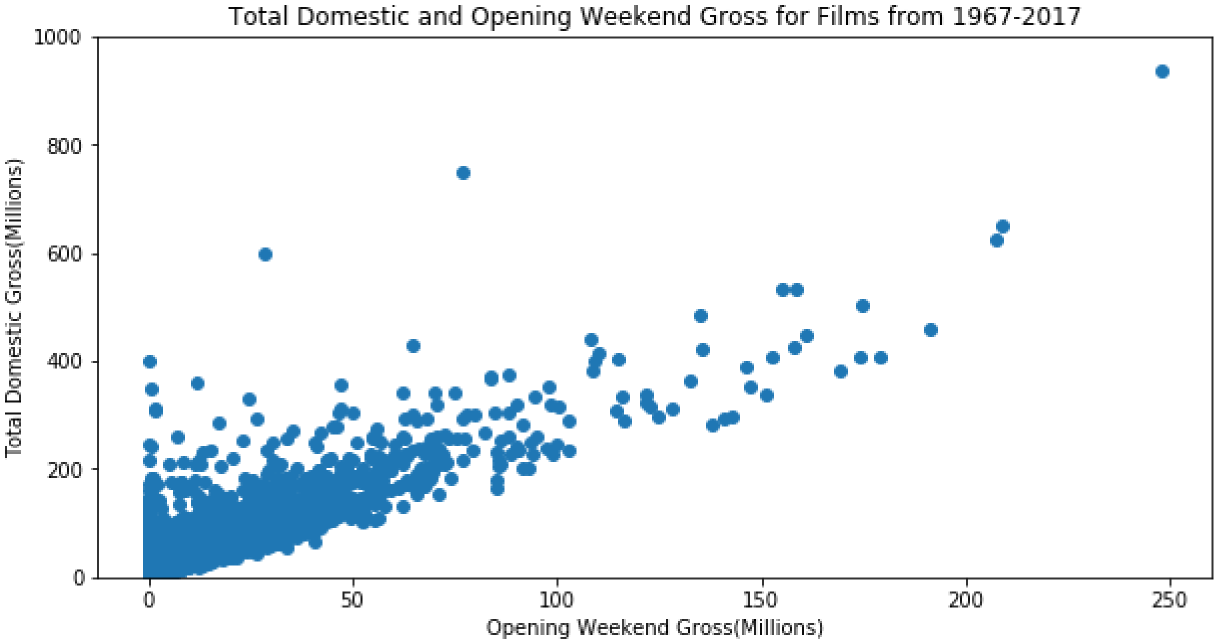
Movie Database (IMDB) ratings, the average score given by users of the website out of 10, alongside the number of votes cast. These features would all prove vital in understanding the relationship between audience and critics reviews and the domestic box office performance of a film. Alongside the critical and audience reception, the OMDB API provided other useful 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 numbers during its initial release is considered, though 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.

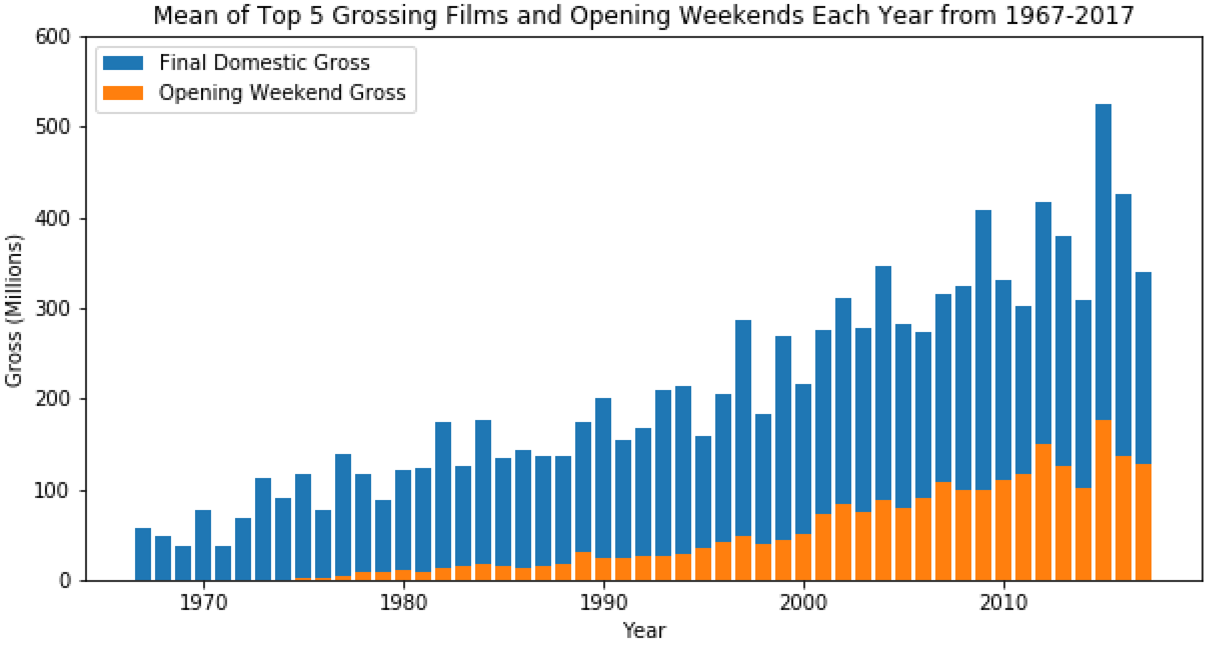
Due to Box Office Mojo and IMDB both being owned by Amazon, all the film's on the Box Office Mojo website had an IMDB ID as well. Using the IMDB ID provided from the base dataset, I requested all the information from the OMDB API for each film of 45,000 films in the original dataset, followed by scraping 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 I exported to a CSV file. Once my movie dataset was created, I explored the data using a few built-in functions 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 serve crucial in the assistance of the predictive model. Opening weekend grosses are rarely substantially impacted by reviews since for word of mouth to build, a large section of the audience needs to see the film first. Opening weekend is a reflection of marketing and promotional efforts. 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 strings, I converted the important metrics to their respectful types, monetary figures were transformed to float objects and so forth.

Exploratory Analysis

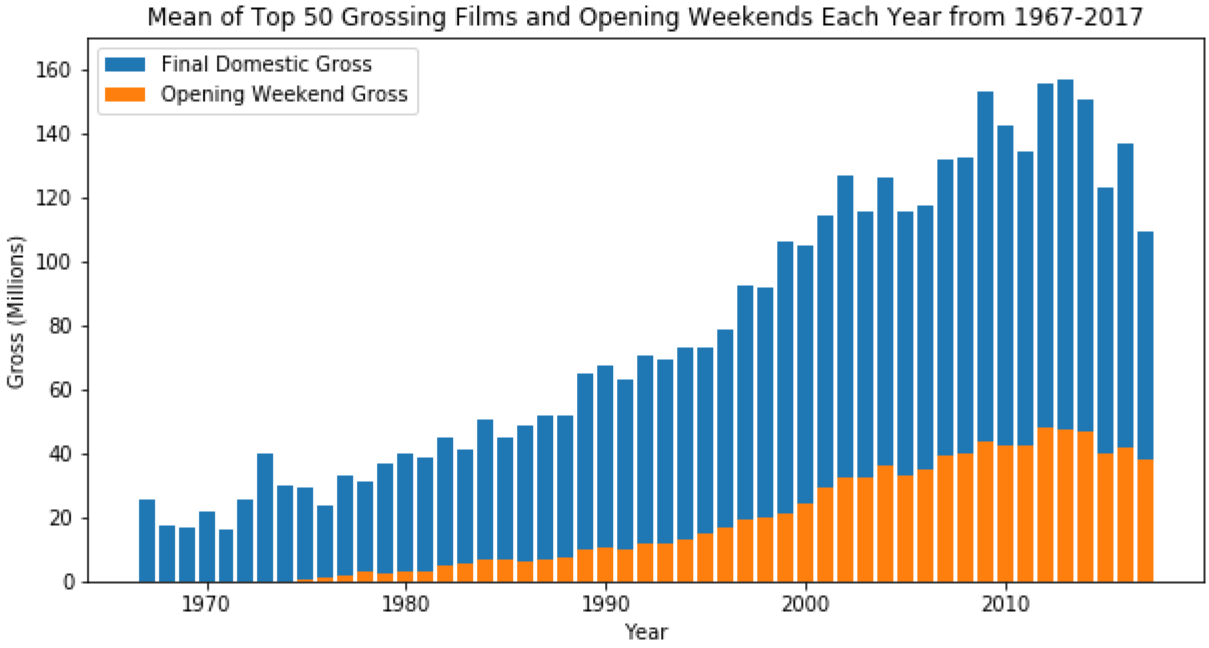
I created a subset of my original movies dataset, isolating only the films that had a domestic box office gross data This new subset contained 12,000 of the original 45,000 films available in the original dataset, the movies without relevant box office data were discarded as they would add little to 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 wanted to focus 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 till 2017. Using matplot lib 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.



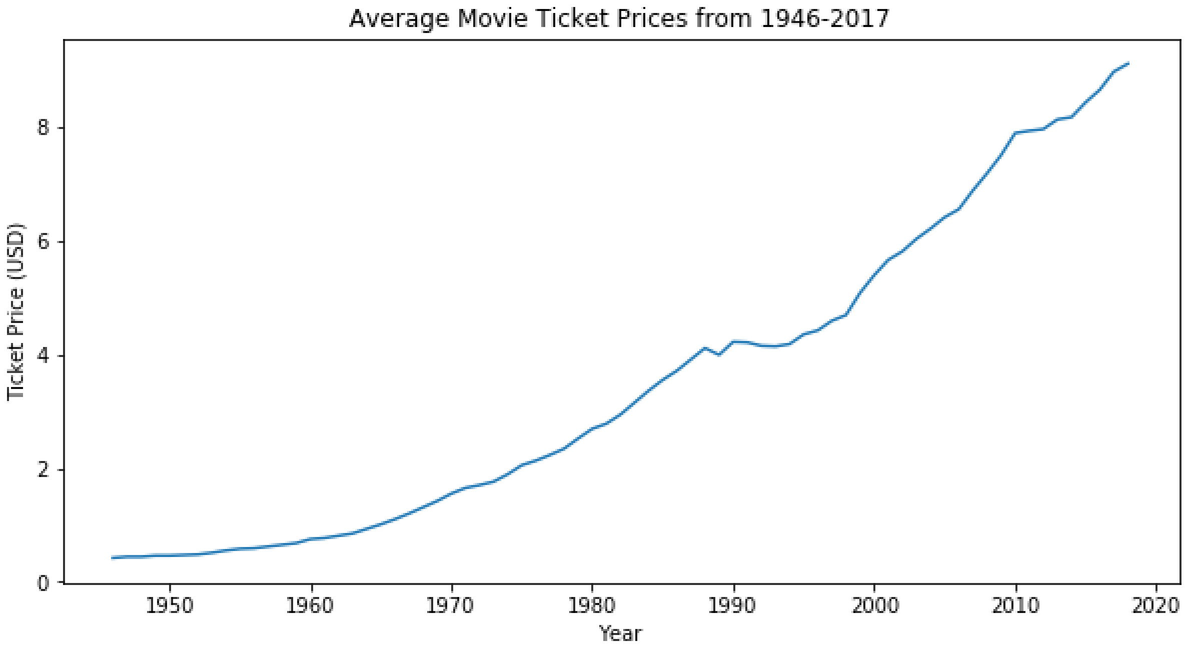
As can be observed from the graph above, there appears to be a relatively proportional relationship between the final domestic and opening weekend grosses for the films in the dataset from 1967 to 2017. Understanding this relationship will be critical for the predictive model and hence looking at how this relationship has changed 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.



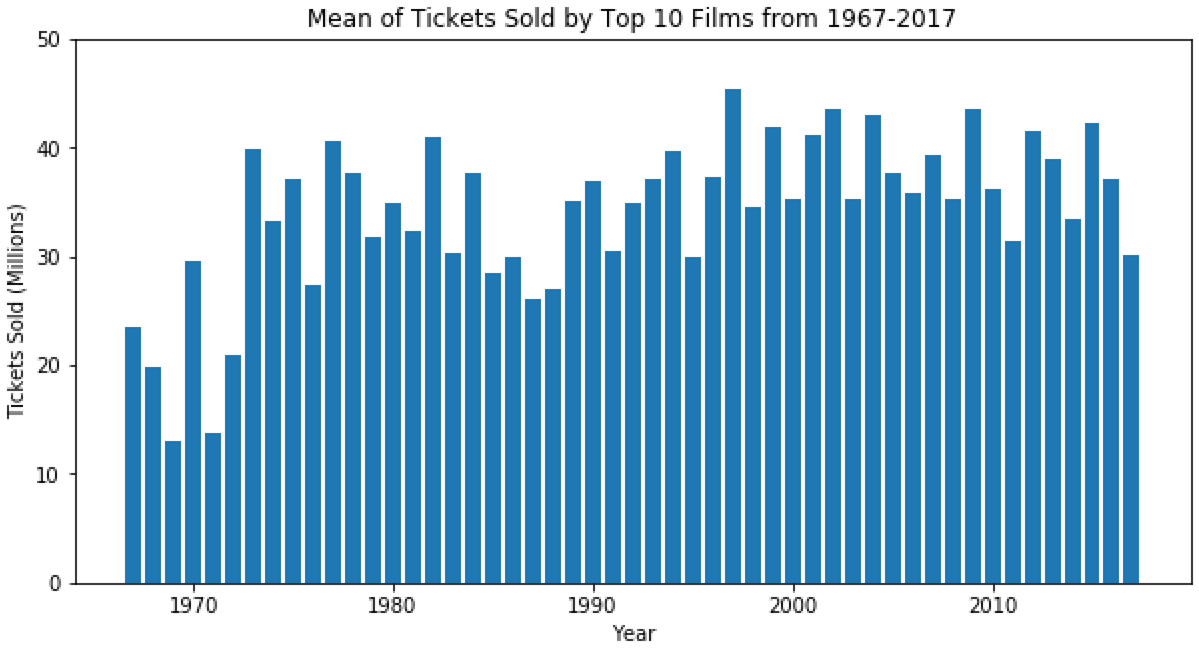
The graph above and below display the trend of both domestic and opening weekend grosses throughout the last forty years. The graph above takes the mean of the top five highest-grossing opening weekends and domestic grosses for each year, this graph represents the trend for films that were very successful 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 of opening weekend grosses, in both graphs growing at a far faster rate than domestic gross of the last forty years. Also, notice that 2015 in the graph above, 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 early 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 year prior to 2017 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 below displaying the mean of top fifty films of that year, baring 2017, 2015 has the lowest mean in ten years.



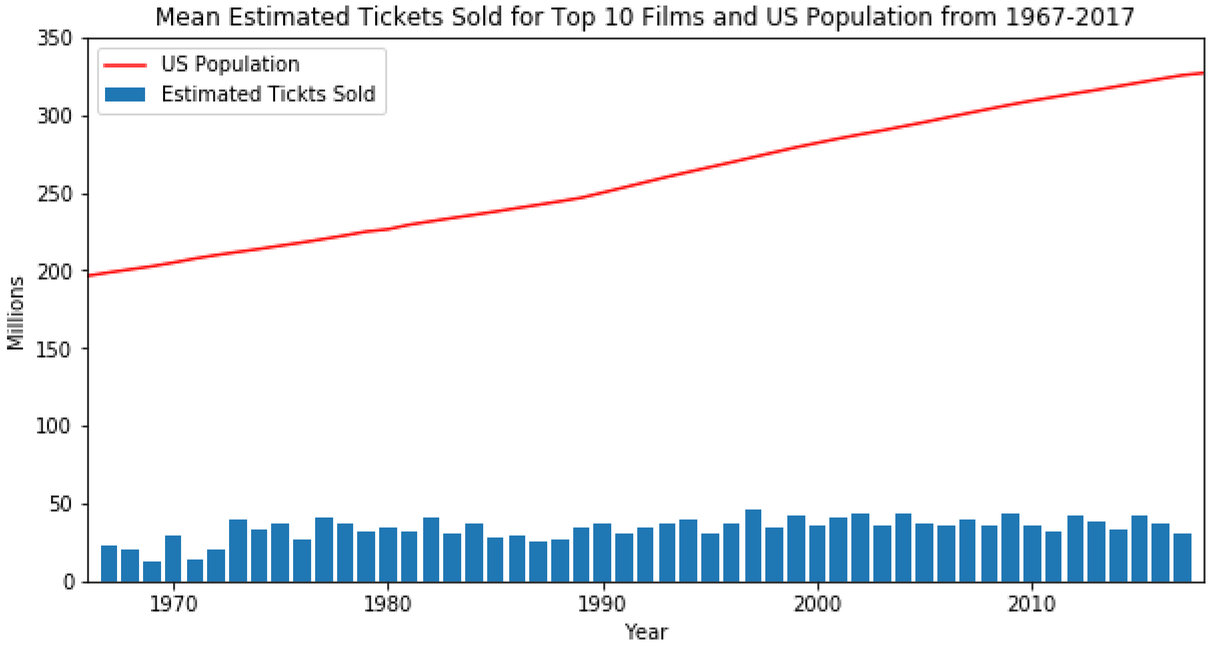
While we see positive growth in both opening weekend and final domestic grosses for films in the last forty years it is important to consider the rise in ticket prices over this period of time. Below is a graph of the national average ticket prices from 1946-2017.



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.

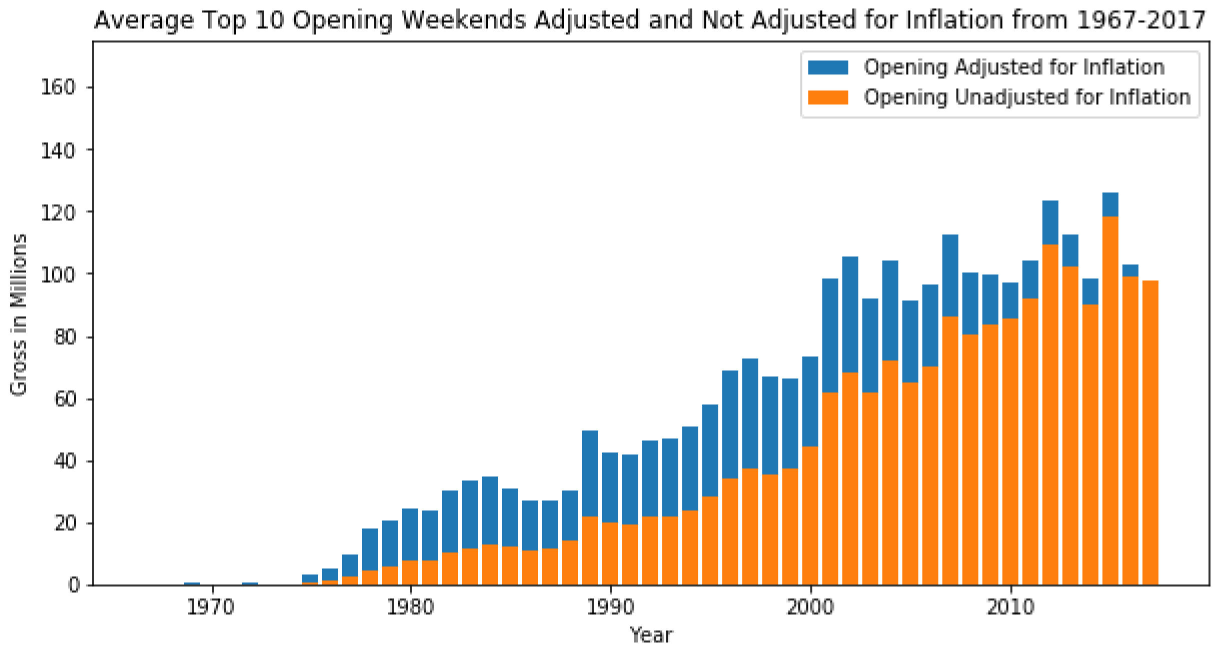


We can observe a relatively consistent number of tickets being sold in this period of forty years, a majority selling over 30 million and under 50 million tickets. There is a larger disparity between years early in the graph, partly due to box office figures for many films in that era not being available.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

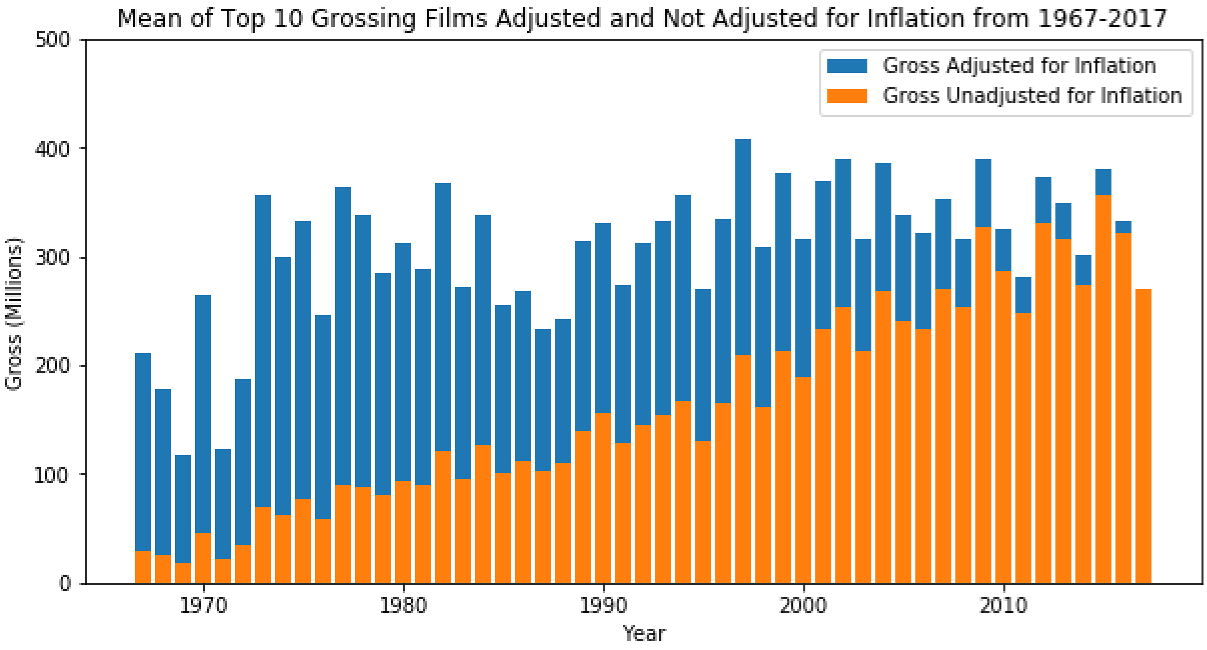


Another point to note for the graph above is that only the population of the United States is shown while tickets sold in Canada are also included in the estimates which makes up approximately ten percent of tickets sold. Where the tickets sold remain consistent as the prior graph suggested the population 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 an eighth of the population. While 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.

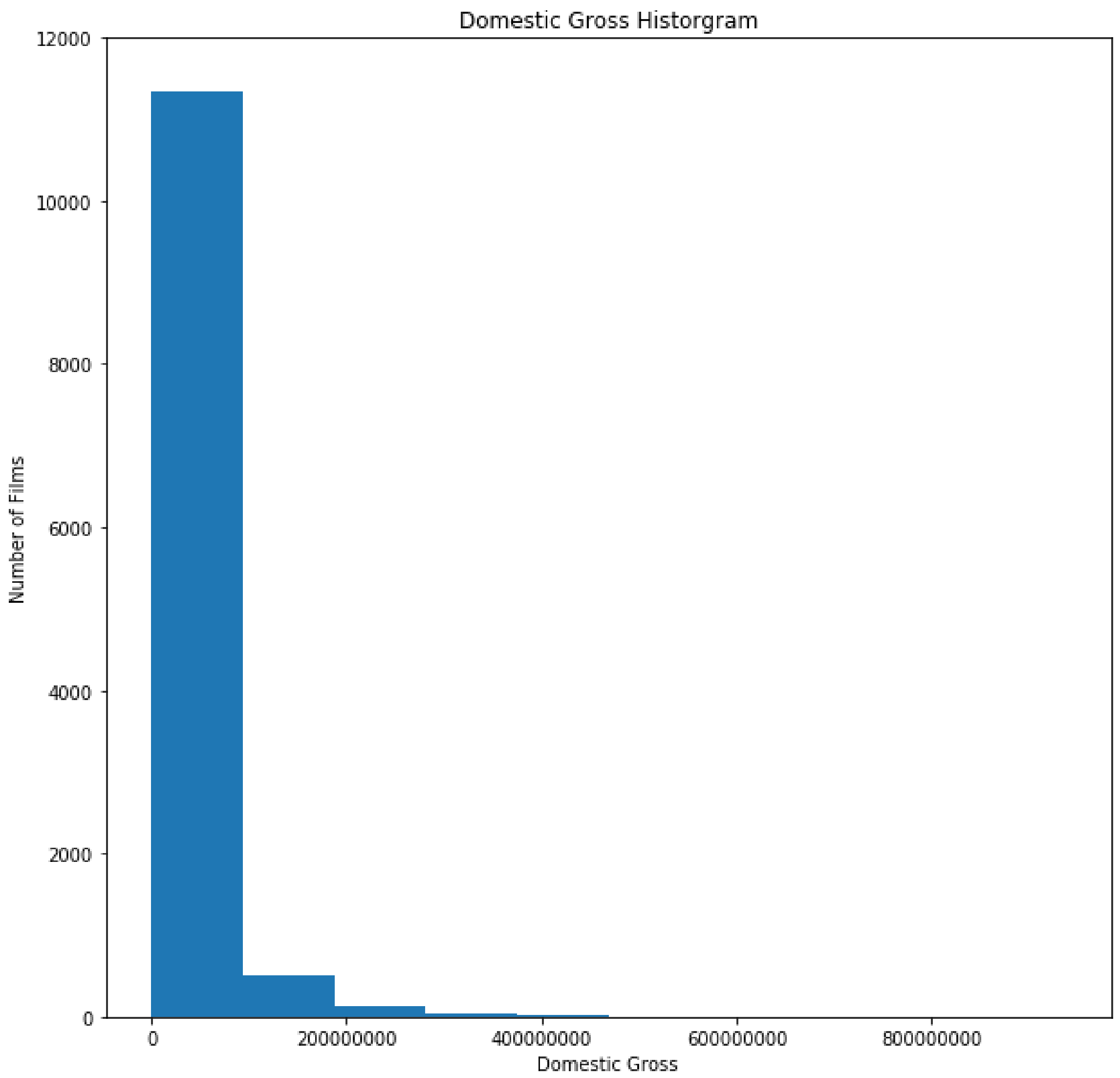


Average Top 10 Opening Weekends Adjusted and Not Adjusted for Inflation from 1967-2017



Both graphs validate our earlier claim, the opening weekend grosses have seen significant growth over the four decades, while the growth of domestic gross can be confirmed to be due to inflation of ticket prices. There exists little opening weekend data for films of the 1970s and 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.

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 shows the histogram for the domestic box office gross from my dataset.

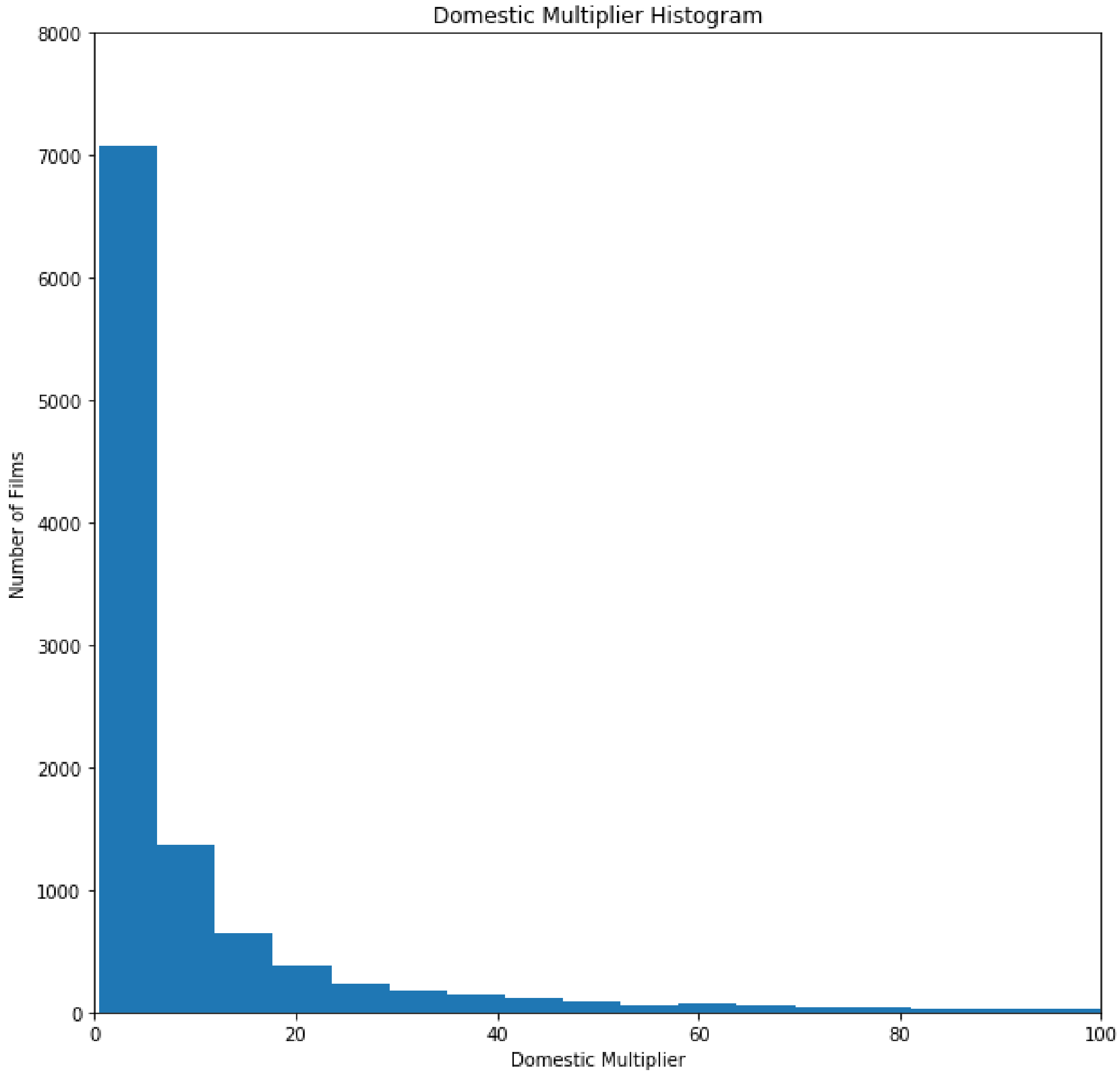


The majority of the domestic gross was under 100 million dollars, with the number of films grossing above 400 million dollars virtually invisible in the histogram above. Alongside the histogram I used the pandas built-in function, describe on my box office dataframe which gave me insights such as the mean of the domestic gross column in my dataframe 21.5 million and the standard deviation 47.2 million, confirming that the data has high variance. 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 tested was, the box office gross of a film and its Rotten Tomatoes score, a measure of its critical response, were independent of each other. To separate positive critical reception from negative reception, two subsets of the box office dataframe, one with a 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 about 20 from the mean Rotten

Tomatoes Score, 59 for all films in the box office dataframe. Each subset dataframe contained around 3,500 films further validating the boundaries used to split the data. Films with mixed reception were ignored as they would not assist in our primary goal of discovering whether positive or negative critical reception indeed 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. 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 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. As discussed earlier 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. 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 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 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.