**Data Wrangling Final Project Report**

**Movies Released in the Past Ten Years in the U.S.**

**By: Lindsey Bolton and Mollie Cox**

1. **Introduction**

Many people believe that movies that generate high box office sales1 will have high critical scores. However, box office sales are technically determined by the public. So, is there a correlation between critic reviews and sales? In this project, we will conduct an analysis of these variables using Metacritic's2 critic review scores and IMDb charts. We plan to use Metacritic’s critic review scores to see if there is a correlation between movie score and box office sales from IMDb3 charts.

1. **Data**

The first data source we used is Metacritic. This website tracks new movie releases and their respective scores, which are further broken down into “user score” and “meta score”. For this dataset, we focused on the “meta score” since it is comprised of critic scores. We scraped the Metacritic website under the “New Releases” page for the past 10 years of released movies using the R package *xml2.* From this website, we took the movie title, the year released, and the critic review score. After scraping 399 pages of this dataset, we were left with 9,576 entries to clean. To begin, we had to remove leading numbers in front of the title of each movie. Next, we had to separate the date and movie rating. For easy analysis, we decided to only use the year of each date, so we took just the year from the dates remaining. Finally, we trimmed any whitespace in all of the columns.

The second data source we used is IMDb charts. This website tracks new movie releases along with their total box office sales in the US. Similar to the last data source, we scraped this website to obtain the movie title, genre, and total box office sales over the past 10 years using the R package *xml2*. After scraping 100 pages, we were left with 6,991 entries to clean. IMDb did not provide an easy way to scrape their data, so we had a lot of cleaning to do. In scraping, the website pulled both “Yearly Gross” and “Total Gross”. Since we only wanted total gross, we had to remove every other entry. We also had to remove other characters from the gross so we would be left with only the dollar amount. Similar to our previous scraping, we had to separate the year and the leading number from the title. While scraping, IMDb provided us with multiple genres assigned to each movie. We decided to work with just the first genre listed, so we had to separate and clean that column as well. Finally, some movies did not have a total gross associated (either due to a lack of information or a lack of gross altogether). We deleted these entries so we were only left with movies we could perform analysis on.

To merge the data, we had to reorder the columns to be in similar order. Then, we horizontally merged the data on the condition that “title” and “date” were identical from both datasets. Any entries that did not match were deleted from the merge, so we would only have entries that we could perform analysis on. This left us with 1,144 entries in our final dataset.

Finally, we performed our analysis on our merged dataset. To complete this, we created 4 different charts for each research question. We utilized a scatterplot to understand correlations, a bar chart to understand the quantitative values assigned to categorical features, and a line chart to understand yearly trends.

*Table 1 Data Dictionary*

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Type | Source | Description |
| title | text | both | Title of the movie. |
| score | numeric | Metacritic | The review scores that critics gave the movie. |
| gross | numeric | IMDb | The total revenue the movie generated. |
| date | date | both | The date that the movie was released. |
| genre | text | IMDb | The category the movie belongs to. |

1. **Analysis**

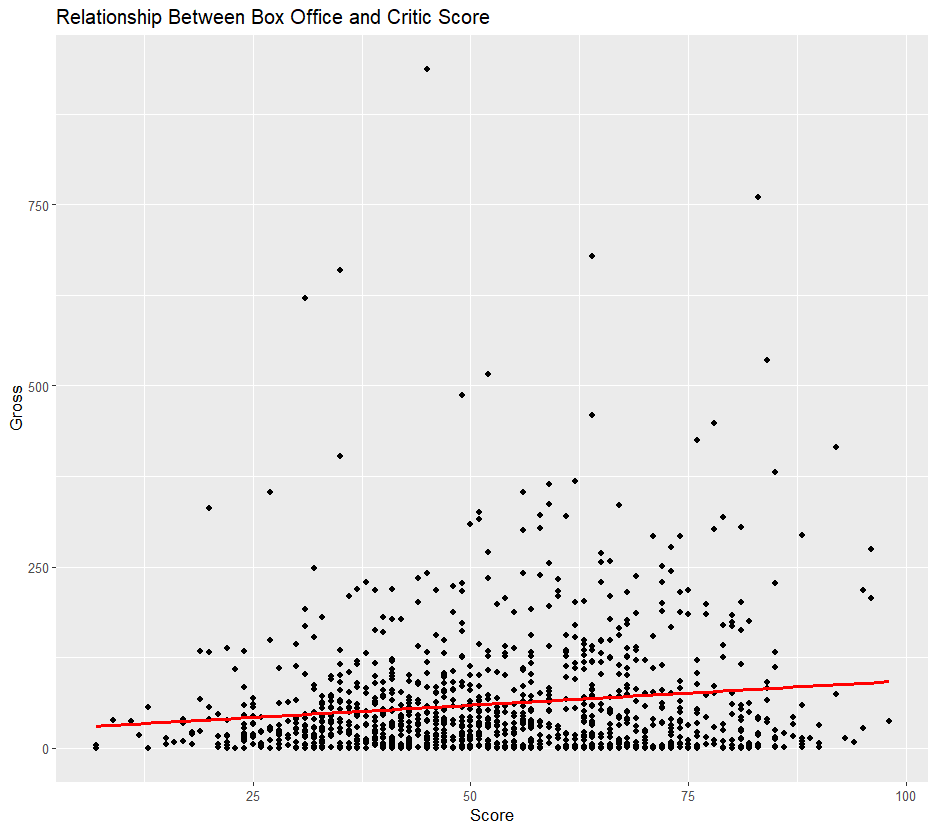
This project aims to understand the relationship between box office sales and critics scores for movies from 2006 to 2022. We are also looking to explore how genre can affect box office sales, as well as the years in which the movie industry had the highest sales total. We will dissect the data to identify patterns that show how genre-specific characteristics impact these metrics. Lastly, we are interested in figuring out which years stood out for the movie industry, and looking at whether there has been a rise or fall in movie popularity. We have explored the following research questions:

**3.1**

Is there a strong relationship between box office sales and critic reviews? Can we use the performance of a movie based on critic reviews to predict box office sales?

There is not a strong correlation between box office sales and critic scores. We created a scatterplot to show the correlation between Box Office Gross Profit and Critic Score, as you can see in *Figure 3.1.* The correlation line shows that there is no strong correlation between these metrics. While critics reviews are usually seen as a predictor of a movie's success, numerous factors contribute to the disconnect between the two metrics. Audiences' preferences and tastes may differ significantly from those of critics, bringing a discrepancy between the critic’s scores and the public’s viewing of movies. Additionally, marketing strategies, star appearances, genre preferences, and the timing of a film’s release can heavily influence box office performance, overshadowing the critic scores. In addition, films with a strong fan base or established franchises thrive at the box office despite negative scores and reviews. Subjectiveness can also play a part in critic scores of the films they are evaluating.

*Figure 3.1 – Box Office Gross vs. Critic Score*



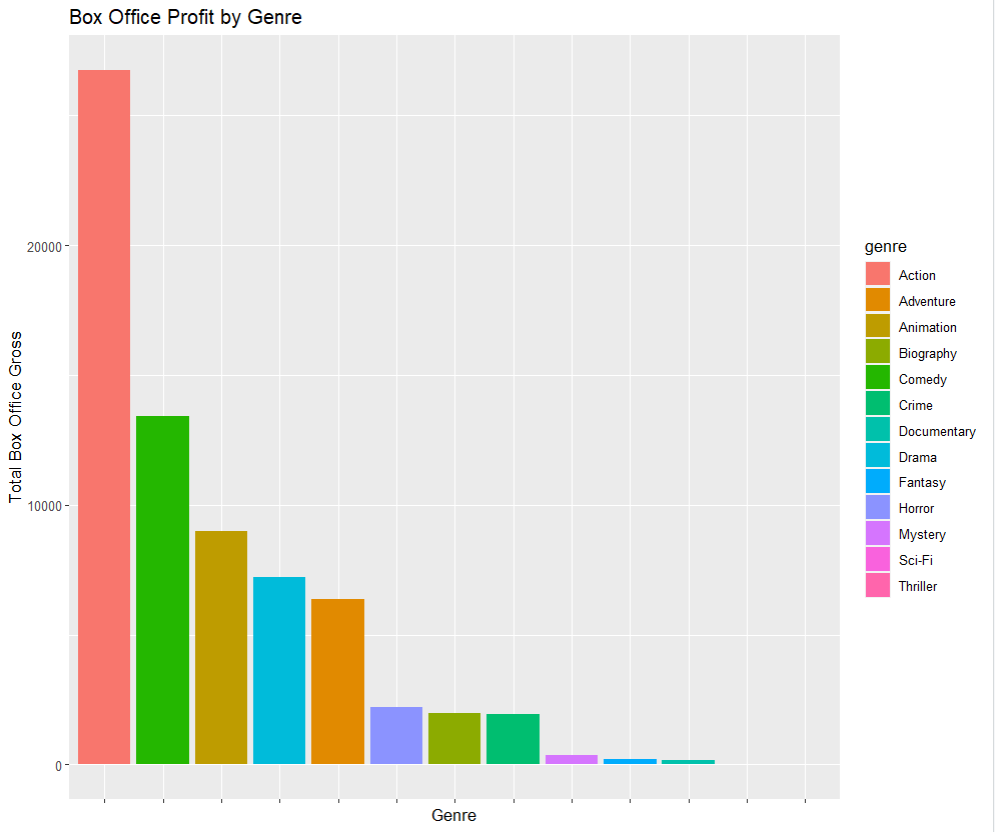
**3.2**

Which genre generates the most box office sales? Do critics and the public tend to feel the same way about certain genres?

In our analysis, we have found that action movies generate the most box office sales. Our graph provides intriguing insights into the dynamics of box office sales across various film genres. It becomes apparent that action movies consistently are the top contender in generating the highest box office sales. This observation aligns with the apparent consensus between critics and the public, who both demonstrate enthusiasm for action-packed films. Factors contributing to the box office success of action movies may include the adrenaline-pumping scenes, high-stakes narratives, and the cinematic experience they offer. The data suggests that audiences gravitate towards the thrilling and visually impactful nature of firms, making them a reliable choice for filmmakers seeking success.

Surprisingly, the analysis of box office sales data points to sci-fi and thriller genres consistently having the lowest box office profit returns. Several factors could contribute to this trend, due to the audiences' opinions. While these genres may have dedicated fan bases, their appeal may be more niche compared to other genres that are widely popular. Thrillers are not seen as family-oriented films and are often R-rated. The success of these movies often relies on critic scores and word-of-mouth, which are difficult to predict.

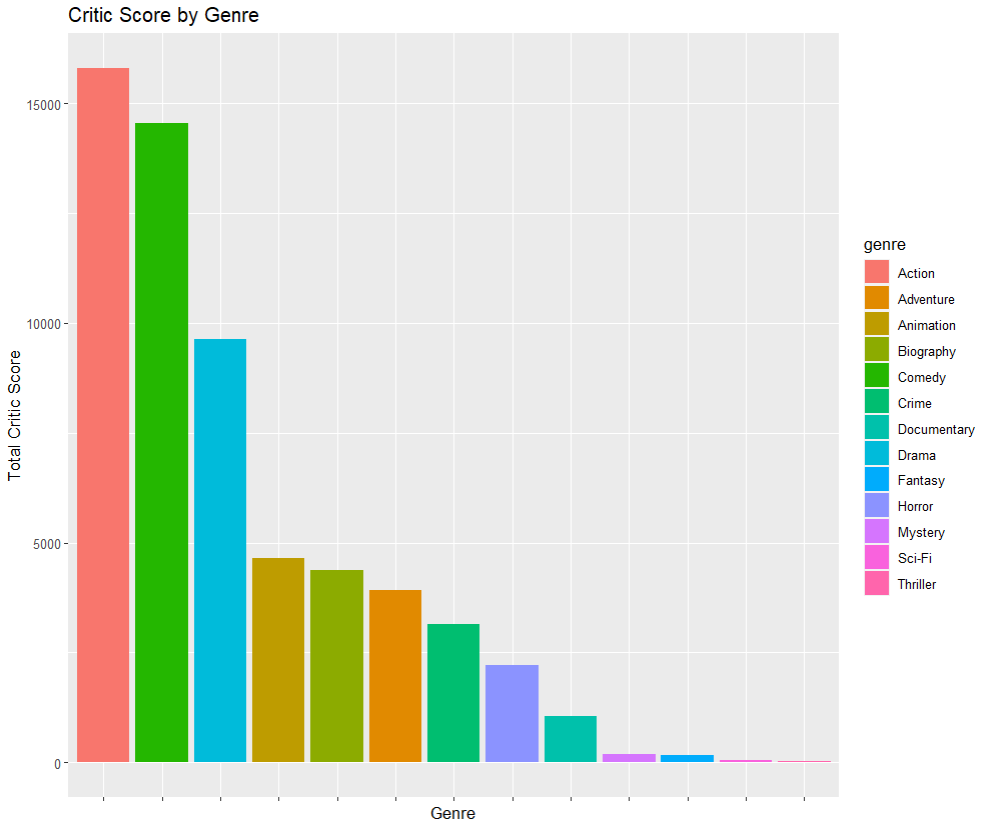
*Figure 3.2 – Box Office Profit by Specific Genre*



Our analysis extends beyond box office sales to explore the relation between critics scores and movie genres. Furthermore, we also analyzed the relationship of critics scores and the genres that have the highest scores. These findings correlate with *Figure 3.2,* indicating that there is a relationship between box office sales by genre and the critics score of those genres. *Figure 3.3* illustrates the connection between critic scores and specific genres. This alignment underscores the influence critics bring in shaping audience perceptions and potentially impacting the viability of movies within certain genres. As said previously, individual tastes and preferences play a role in the correlations, but the data suggests that critic scores can act as a predictive factor for a genre’s overall performance at the box office. This offers valuable insights for filmmakers seeking to navigate cinematic success.

Thrillers and sci-fi movies continue the relationship as seen in *Figure 3.2* into *Figure 3.3* being the lowest rated movie genres by critics. This can confirm the predictions made by our earlier analysis.

*Figure 3.3 - Critic Score by Specific Genre*



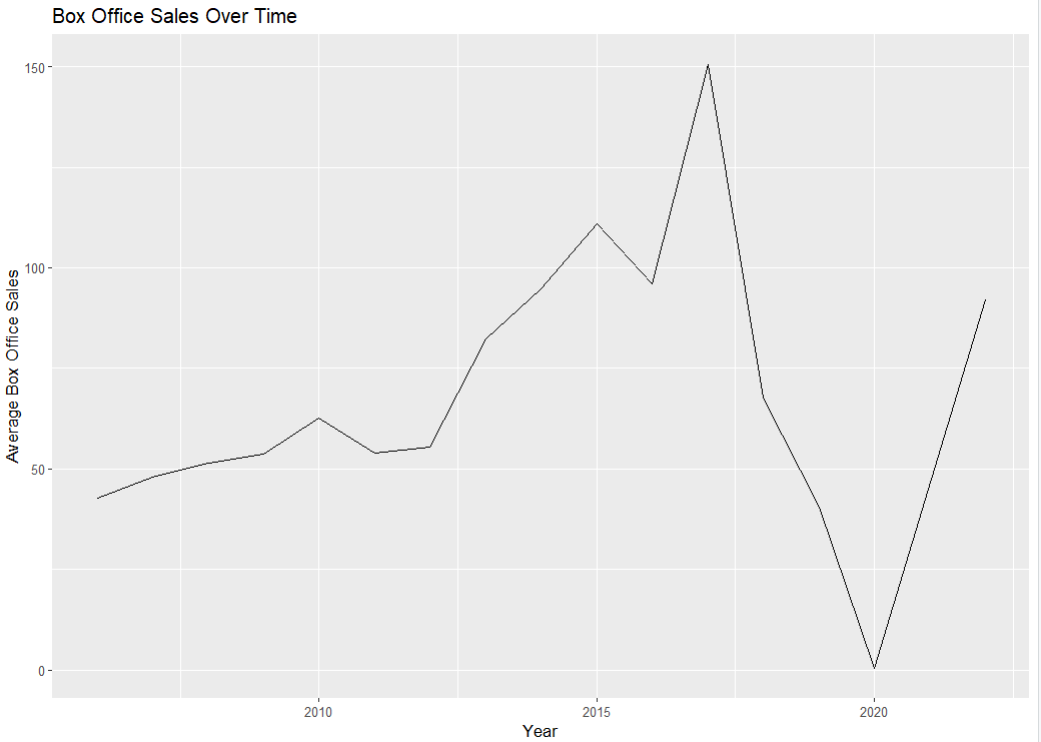
**3.3**

What years have seen the most box office sales? Have movies increased or decreased in popularity, and is there an explanation for these trends?

In 2017, the movie industry experienced a peak in box office sales, marking it as a standout year for cinematic success. As pictured in *Figure 3.4,* both 2018 and 2019 witnessed a decline in sales, indicating a temporary dip in popularity. However, the fortunes of the industry took a positive turn after, with a noticeable increase in sales. Several factors may have contributed to this fluctuation. The decline in 2018 and 2019 could be attributed to a combination of popular markets, changing consumer preferences, and also a shortage of blockbuster releases that capture widespread attention. The later uptick in sales may reflect the industry's adaptive response, possibly driven by strategic timing, diverse content offerings, and innovative marketing approaches. This dynamic pattern underscores the resilience of the movie industry, which, despite facing periodic challenges, demonstrates the capacity to reinvent itself and recapture audience interest over time.

An important point to address is the significant impact of the COVID-19 pandemic in early 2020. The global health crisis halted the release of new movies and made it impossible to produce any films during this time. This also accounts for low box office sales since in-person viewing was restricted. This certainly influenced the initial climb of box office sales starting in late 2019. As restrictions eased and the industry adapted to new requirements, there was a rise in box office sales from 2020 to 2022.

*Figure 3.4 - Box Office Sales over Time*



1. **Conclusion**
2. *Is there a strong relationship between box office sales and critic reviews? Can we use the performance of a movie based on critic reviews to predict box office sales?*

There is no correlation between box office sales and Metacritic reviews. Therefore, we cannot use the scoring of a movie to predict box office sales.

1. *Which genre generates the most box office sales? Do critics and the public tend to feel the same way about certain genres?*

Action movies generate the most box office sales. Critics and the public tend to feel the same way about certain genres, as action is also the highest rated amongst. Critics

1. *What years have seen the most box office sales? Have movies increased or decreased in popularity, and is there an explanation for these trends?*

2017 had the highest box office sales. Movies decreased in sales in 2018 and 2019, indicating a decline in popularity. However, sales increased after the global pandemic in 2020.

After completing the analysis of our data, we have found numerous limitations that are important to consider moving forward.

* ***Lack of data:*** After merging our datasets, we found that no entries included the year 2021. This means that our analysis is not as detailed as we would have preferred and is missing key data. In the future, we would work to scrape more websites to ensure we had all the metrics necessary to perform an adequate analysis.
* ***Inconsistencies:*** Because our two scrapings had different data, not every item was kept for our analysis. This means that some important movies were removed from our dataset. This resulted in inconsistencies with our data and our common knowledge. For example, our analysis found that 2019 had one of the lowest box office sales. However, we know that Avengers: Endgame was released that year and continues to hold the highest box office (without inflation). Upon examining our dataset, we found that Avengers: Endgame was not included for analysis. In the future, we would potentially scrape other websites to ensure top earning movies were included.

There are many other features that could be explored in future iterations of this project. For example, we could look at whether movie ratings (i.e. PG, PG-13, R) have any impact on box office sales or critic rating. Also, implementing a code to update and compare box office sales adjusted for inflation could provide interesting insights.