Movie's Data (2014-2021) Analysis Report

Satyanarayana Vinay Achanta (A02395874)
Aashay Maheshwarkar (A02397403)
Megh KC (A02373436)

Introduction:

The movie industry is one of the most profitable industries in the world. In this project, we aim to gain insights into the trends and characteristics of the movie industry by analyzing a dataset of movies from 2014 to 2021 which we scraped from the IMDb website with Beautiful Soup. Our analysis techniques include trend analysis, correlation analysis, and statistical tests.

Our dataset contains information on movie titles, release years, genres, ratings, meta scores, box office revenues, votes, directors, and stars. We used this dataset to answer five key questions: 1) What are the trends in gross revenue for movies from 2014 to 2021? 2) Is there a relationship between IMDb votes and box office revenue? 3) What is the proportion of movies by genre? 4) Is there a relationship between gross revenue and IMDb votes? 5) Are there significant differences between director and movie ratings?

The purpose of our analysis is to provide valuable insights to stakeholders in the movie industry, such as movie producers, directors, and investors. By understanding the trends and characteristics of the movie industry, they can make better decisions and investments. For example, our analysis of gross revenue trends can help producers and investors to identify the most profitable genres and years. Our correlation analysis can help them to understand the relationship between IMDb votes and box office revenue, which can inform marketing and distribution strategies.

GitHub Link
Presentation Slides Link

Dataset:

The dataset contains over 100,000 movies and includes important attributes such as movie name, release year, rating, meta score, votes, gross revenue, genre, director, and stars.

The dataset was obtained by scraping the IMDb website and merging the data from seven separate CSV files. Gross earnings were used to account for inflation measurements. The analysis includes five main sections: Trend Analysis on Gross Revenue of Movies Trended from 2014 to 2021, Exploring the Relationship Between IMDb Votes and Box Office Revenue, Proportion of Movies by Genre: A Pie Chart, Examining the Relationship Between Gross Revenue and IMDb Votes, and Comparing Director and Movie Ratings: A T-Test Analysis. We merged our scraped data with a CSV file which we found from this link, as a CSV.

name	The Movie name in the IMDb website
year	The year of release of the movie
gross	The gross earnings made by a specific movie
genre	The information about the genre of a movie
Amount	Variable on with the inflation adjustment factor is calculated
Rating	The Movie's IMDb rating
Director	The Directors of a particular movie

To collect data for our movie analysis project, we used Beautiful Soup to scrape data from the IMDb website. We provided the URL for the years we wanted to scrape and set the number of movies using the 'target' variable. In a loop, we requested each page, parsed the HTML source code

with Beautiful Soup, and extract information such as movie name, release year, gross, genre, ratings, meta score, votes, stars, and directors. We found all the movie listings on the page using 'findAll' and extracted their attributes using the 'extractcolumns' function. We checked if we had scraped the desired number of movies and continued until we did. Finally, we converted the scraped data into a pandas DataFrame and saved it to a CSV file called 'Movie.csv'.
The above columns(in the table) were extensively used in our analysis.

Analysis Technique:

After picking some factors we were interested in we removed rows with empty cells. We cleaned all the entries to get rid of punctuation and just have raw data. The analyses we made were:

1. Analysis-1: Gross revenues of movies from 2014 to 2021
    - Total gross income: we gathered the gross income for each movie and converted them into present dollar values using the inflation rate. We chose to visualize this with a line chart, as our aim here was to see the temporal trends compared to the gross earnings.

2. Analysis-2: Exploring the Relationship Between IMDb Votes and Box Office Revenue
    - We have plotted a regression plot, as we wanted to see a correlation between the number of votes on the IMDb website and the gross revenue earned by the movie. The votes earned by a movie can be seen as a factor in its popularity.
    - Pearson correlation test to determine the confidence and strength of correlation.

3. Analysis-3: Examining the Relationship Between Gross Revenue and IMDb ratings
    - Looking for a correlation between the movie quality and the revenue earned, the quality is referenced here as the IMDb rating for a movie.
    - Plotted a Regression plot to see a correlation between the fields, to get a sense of the strength and confidence in our results, we did a Pearson correlation test.

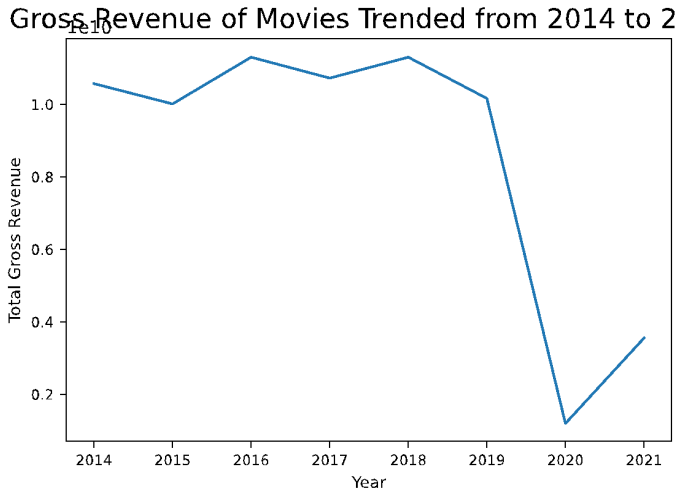
4. Analysis-4: Proportion of Movies by Genre: A Pie Chart
    - There are different genres and all cannot be included. We included genres and would like to see how much each contributes to the movie industry. We chose overall revenue collection data. A pie chart was chosen here, because we wanted to get the ‘share’ of each genre in the market, making it a suitable visual aid.

5. Analysis-5: Comparing Director and Movie Ratings: A t-Test Analysis
    - The distribution of movie ratings for movies directed by different directors and perform a t-test to see if there is a significant difference in the average movie rating for movies directed by different directors
    - We drew a Kernel Density Estimation(KDE) plot here, to visually see the differences in the mean, standard deviation, and see the distribution as a density function, for the two directors in consideration.

Results:

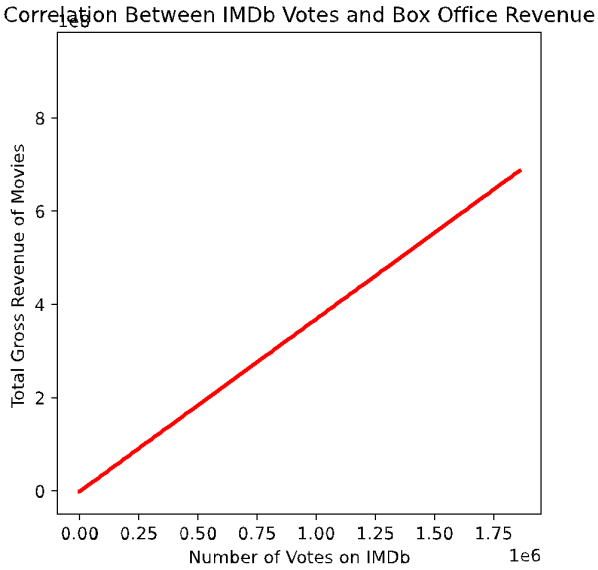
Analysis-1: Trend Analysis on Gross Revenue of Movies Trended from 2014 to 2021

We tried to check the relation between getting votes and their box office collection (i.e. which type of movie gets higher votes?) and found that the gross revenue collection (inflation-adjusted to present value) is unpredictable but remains almost the same every year. We saw an unusual drop in movie revenue in the year 2020 which could be because of the covid pandemic. After that, the movie industry is trying to take its shape.



Analysis-2: Exploring the Relationship Between IMDb Votes and Box Office Revenue.we can see Herding behavior here, people watch a movie based on the popularity.

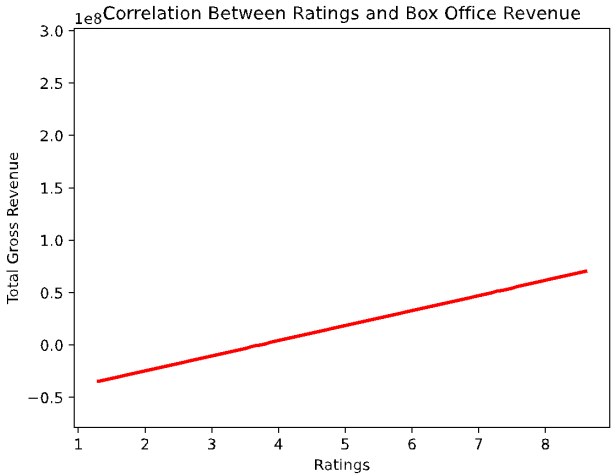
The scatter plot for the #votes for movies and their gross collection could be interesting to see and we did that by plotting the scatter plot. We also make these two variables and put them into Pearson correlation tests. The correlation test showed us a strong correlation between getting votes and making revenue. The probability value of 0 and test statistic value of 0.73 tells us that the commercial movies would get the higher number of votes.




```
(r,p) = stats.pearsonr(df_merged_c.Total_Votes, df_merged_c.Gross_USA)
print('r =', r, 'p =', p)
```

```
r = 0.7302914541360278 p = 0.0
```

Analysis-3: Examining the Relationship Between Gross Revenue and IMDb ratings



We had to answer one burning question, “are higher rated movies appreciated as much as they should be?”. The answer was an astounding yes!.

Pearson Test

```
(r,p) = stats.pearsonr(gross_ratings_data.Rating_Score, gross_ratings_data.Gross_USA)
print('r =', r, 'p =', p)
```

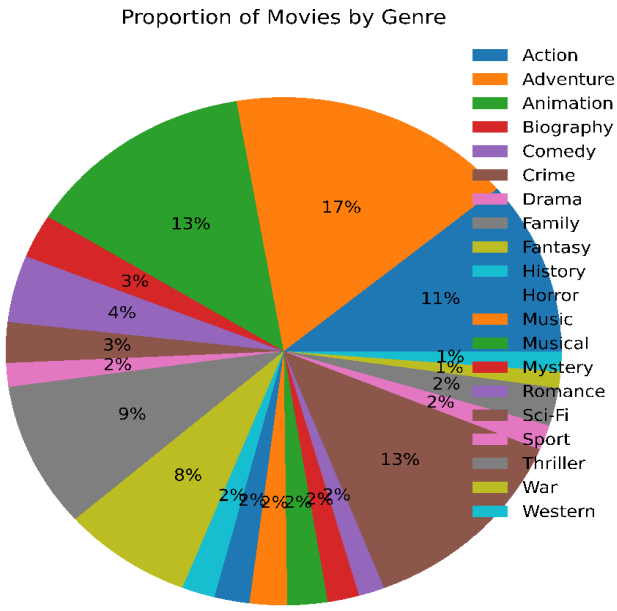
```
r = 0.5636157269115796 p = 1.5126414675906165e-06
```

The correlation plot for the IMDb ratings for movies and their gross collection showed us a moderate positive correlation between ratings and making revenues. The probability value of 0 and test statistic value of 0.56 tells us that the commercial movies would get higher IMDb ratings.

Analysis-4: Proportion of Movies by Genre: A Pie Chart

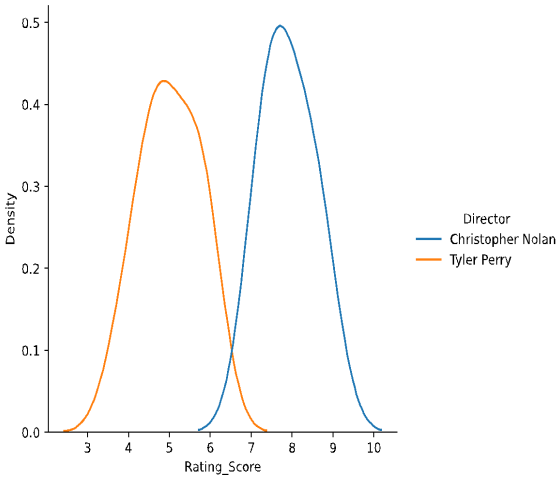
The adventure movie genre came to be the most commercial. People love adventure movies as this genre shares most of the box office collections in the movie industry. Animation movies are the second most watched. The producers can look at that and focus on the most money-making genres in their future movies.

From this analysis we can safely say that to get launched into the movie industry, an Adventure movie would be the best bet, as a lot of them are made and would be a safe bet.



Analysis-5: Comparing Director and Movie Ratings: A t-Test Analysis

We conducted a t-test to see if there is a significant difference in the average movie rating for movies directed by different directors. We took two directors and the movies they directed or were involved in. The mean and standard deviation for movie ratings for Tyler Perry is 5.01 & 0.72 whereas Christopher Nolan has 7.90 & 0.65 respectively. Our hypothesis would be their impact on the movie rating is significantly different. T-test showed a t-test statistic value of -5.90 with the probability being near zero. Hence, the mean movie ratings for both directors are significantly different and that resembles our hypothesis.




```
'Tyler Perry mean: 5.014285714285714'

'Christopher Nolan mean: 7.8999999999999995'

'Tyler Perry sd: 0.7244045173533258'

'Christopher Nolan sd: 0.6557438524302001'

Ttest_indResult(statistic=-5.907628227625489, pvalue=0.00035865000284283156)
```

Technical:

Like the last project, we did data exploration and preprocessing, checked for features we would like to compare and analyze, and removed missing values, data exploration, etc. Also, we have added a function to clear the strings present in the year column. Changed columns like votes, rating, meta score, year, and gross into int as they were in

float before. We used the pandas' package in Python to load all the necessary files (listed in the dataset section). Afterward, we used the group by function to group the data for each analysis and various aggregation functions (sum, mean, max, count) as needed. We did end up merging the two tables provided.

Producing numerical data was a little difficult. We needed to use aggregation to get the data we wanted. Cleaning the cell entries took some time too. The t-test is effective when trying to find if distributions are the same or similar. Beautiful Soup is used to extract the data from the IMDb website and all the information on how we extracted the data is mentioned in the Dataset section. The genres and directors' fields were essentially multivalued attributes, to plot the pie charts, we iterated through the records, and for each genre, we created duplicate records to get a sense of all genres. We have also amalgamated our findings with inflation data, and adjusted our gross amounts with that.

Conclusion of the Story (*THE END*)

Based on the analysis conducted on the movie dataset, it can be concluded that the movie industry has remained fairly consistent in terms of gross revenue collection over the years, with an unusual drop in revenue observed in 2020 due to the COVID-19 pandemic. The correlation tests conducted between votes and box office revenue, as well as between IMDb ratings and box office revenue, showed a strong and moderate positive correlation respectively. Adventure movies were found to be the most commercially successful genre, followed by animation movies. Producers looking to make money in the movie industry could benefit from focusing on these genres. The t-test conducted on movie ratings directed by Tyler Perry and Christopher Nolan showed a significant difference in the average movie ratings for movies directed by both directors, with Christopher Nolan's movies having a higher average rating. Overall, the analysis suggests that adventure movies with high IMDb ratings and talented directors could be the key to success in the movie industry.