Hyperiondev

TASK

Exploratory Data Analysis on the Movies Data Set

Visit our website

Introduction

Summary of the data set

The 'Movies.csv' file is a comma-separated values file that contains information about movies. Each row in the file represents a single movie, and the columns of the file contain various pieces of information about the movie, such as the title, release year and runtime.

There is a total of unique 20No. data columns and 4803No. uncleansed rows.

DATA CLEANING & MISSING DATA

The following methods have been used in cleansing the dataset, where appropriate visualisations are included.

1) Head() – To show data rows and columns, gaining an appreciation of value data types.
2) Info() – To show a summary of the dataframe, including the number of rows and columns, the column names, and datatypes.
3) Isnull() and Drop() – Knowing where any null values are and what our analysis will include allows for the removal of unnecessary data columns. Removing unnecessary data columns simplifies the dataframe. The following data 8No. columns were removed:

    a. Keywords – Not needed.
    b. Homepage – Not needed & significant percentage of null values.
    c. Status – Not needed.
    d. Tagline – Not needed & significant percentage of null values.
    e. Original Language– Not needed.
    f. Overview – Not needed.
    g. Production Companies– Not needed.
    h. Original Title – Not needed.

4) Drop_duplicates() – Given that each row is a separate film, we do not want to duplicate data that may skew any assessment. Duplicate rows were removed. For this assessment, the dataset included no duplicates.
5) Missingno.matrix() visualisation – From observation, some of the rows have zero values which implies their values have not been recorded or some information is missing. '0' values were changed to null values using replace().

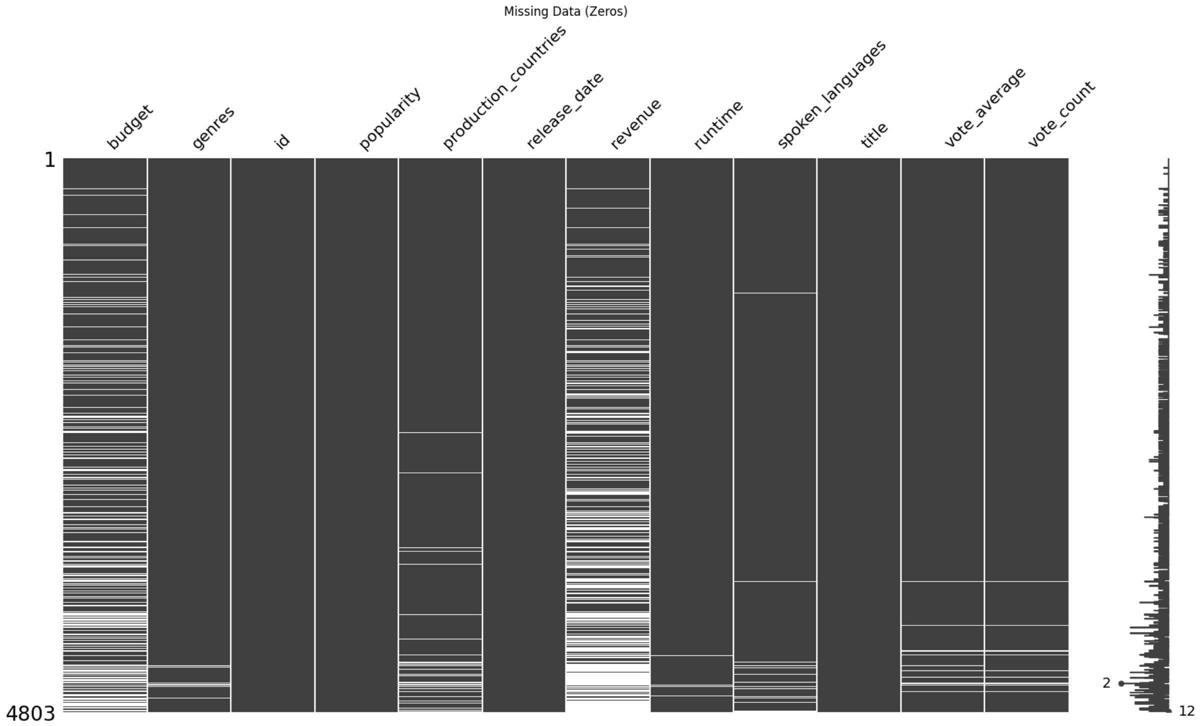


Figure 1 : Missingno Matrix (null values)

6) Dropna() – rows containing converted ‘0’ values to null values removed from dataframe. Total number of rows dropped 1594No., leaving 3209No. complete rows. The missing data (33% of dataset) is considered an example of Missing Completely at Random (MCAR) as missing data is felt to be independent to any other variables in the dataset. The missing data was overcome by deletion, not imputation. It may have been possible to use researching individual films or regression imputation but for this assessment it was felt 33% has too large a portion to comfortably impute at this stage.
7) To_datetime() – Conversion of given dataset release date values to python programme language to represent time and data information. Data was parsed to its own data column. The year information was then extracted and assigned to its own data column.
8) Astype() – To allow for full assessment, both budget and revenue data values cast from object class to integer.
9) JSON Formatted Values – JSON formatted values flattened to make access to data easier. Explode() used to separate flattened list to individual index genre values on duplicate title rows in copied movies dataframe (separate from main dataframe).
10) Minmax scaling() – Due to differences in magnitude and distribution, continuous data columns have been scaled to compare data with different magnitudes and ranges so that one feature does not dominate the other feature i.e., revenue and budget. These scaled data columns are added to the main dataframe.

DATA STORIES AND VISUALISATIONS

Exploratory data analysis has been used to explore various aspects of the film industry such as monetary performance, genres, and popularity. Different visualisations have been used to easier understand and interpret what the data is representing.

Assumptions

1) The given data set contains no unit information, therefore for assessment comments and visualisations presented will be unitless. A higher score value is assumed as being good or bad concerning each data column title.

Patterns and Trends

1) Describe() & Data Distribution with Histograms – To show basic statistics of the numerical columns in the dataframe and review frequency distribution. From this we can review which numerical data is being included and look at the shape of the data included in the dataset.

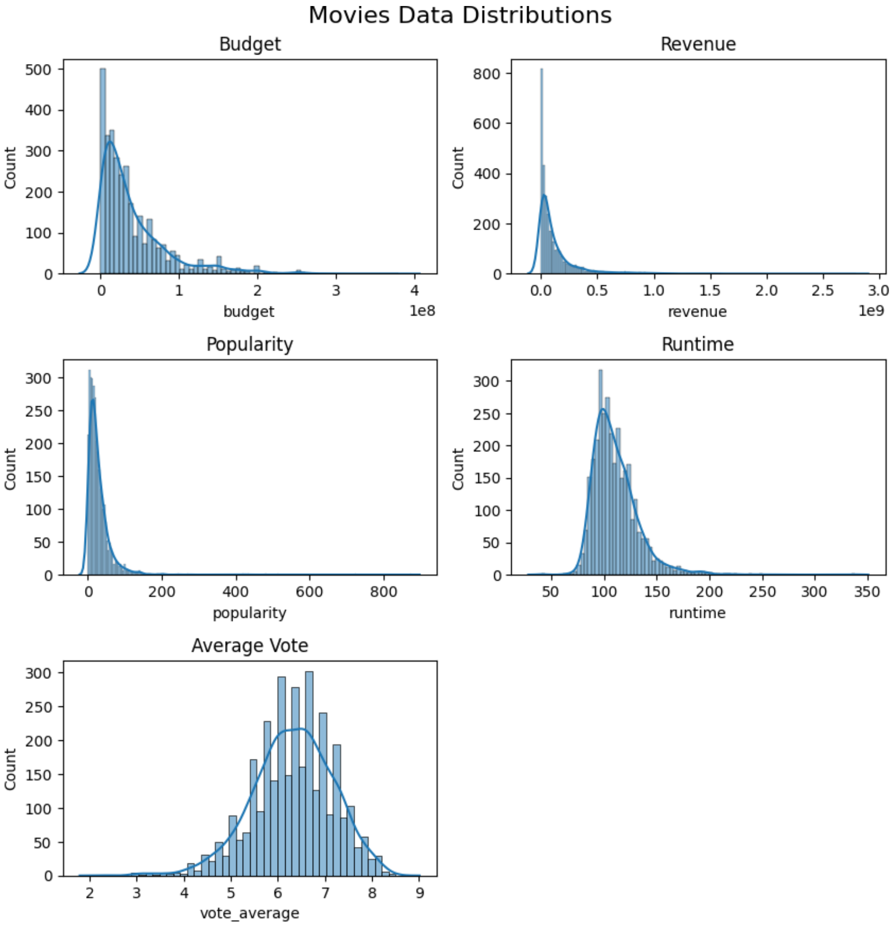


Figure 2: Seaborn Movie Features Histplots Subplots

- The positively skewed distribution of budget, revenue, popularity, and runtime suggests that most movies have lower values for these features, with fewer movies having higher values. This can be interpreted as most movies having a lower budget, revenue, popularity, and runtime, and it is less common or normal to have movies with higher feature values.
- Vote average appears normally distribution but with a longer left tail, with most data at the bells centre with fewer data points at the extreme left and right sides. This tells us that in terms of vote average, scoring is independent and less likely to be biased. Given the longer left tail, voters are slightly more favourable of a scoring above 5 (scoring =0-10, mean 6.3, median 6.3). This suggests that other factors beyond budget, revenue, popularity, and runtime are also important in determining a movie's quality or voter's tastes.

2) Top 10 and bottom 10 Profitable Movies by Scaled Budget and Revenue.

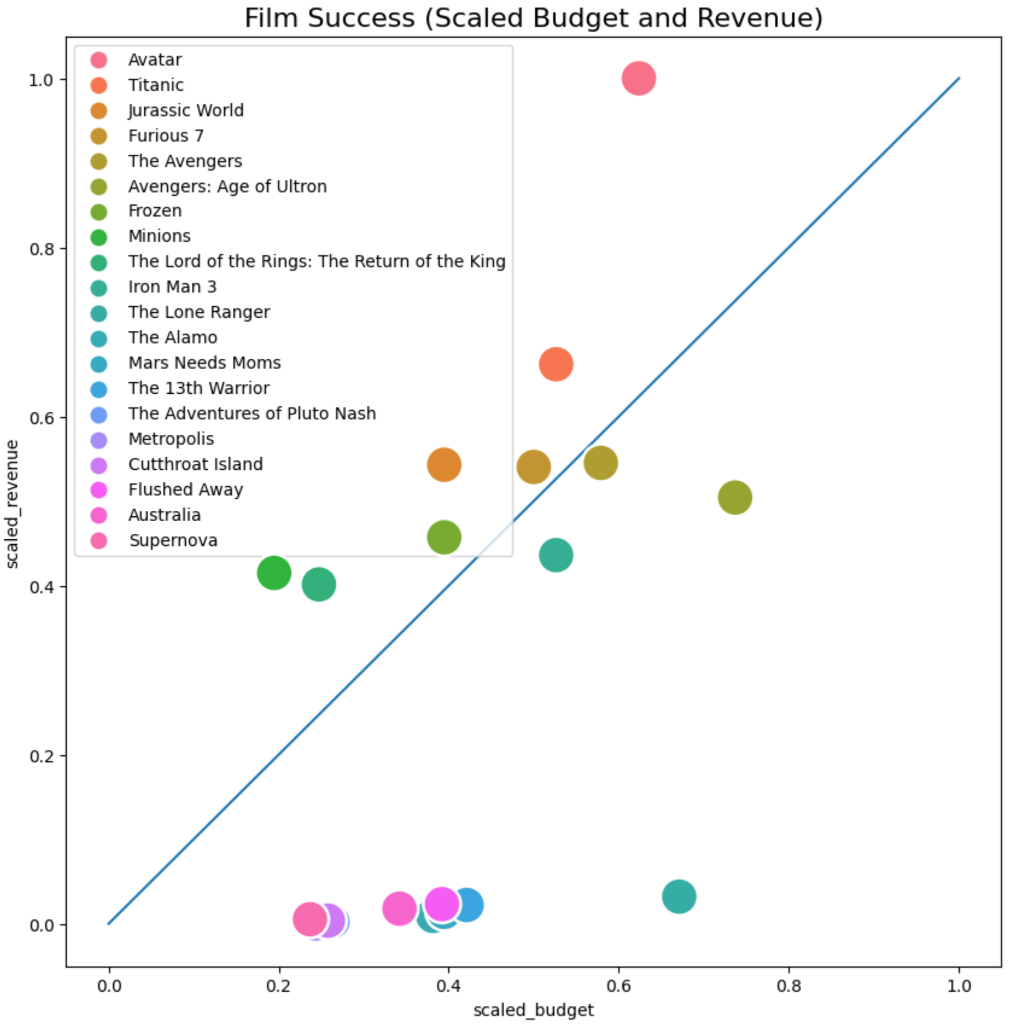


Figure 3: Seaborn Movies Scatterplot (Diagonal Line – 1:1 Scaled Budget/Revenue Ratio)

- From the calculation of profitability (revenue – budget) and concatenation of top and bottom 10No. movies by profit, more expensive movies are not necessarily the highest by revenue, similarly the less expensive films are not necessarily the lowest by revenue.
- Note, the magnitudes between budget and revenue were much different, with mean revenues being x3 greater than mean budgets. When comparing scaled budgets and revenues we begin to see a different story about how successful a film is regarding its given budget and final revenue, so a film with a high revenue may not necessarily be the most profitable.
- Provided with a division line showing a 1:1 scaled budget to revenue ratio, films above that line could be regarded as being more successful than a film with similar budget or revenue value sitting at over below the line or to the right. From the visualisation above we see the following films as being the most successful per unit of budget spent. Surprisingly, despite making more revenue, The Avengers (2012) vs. Frozen (2013) is a good example of this.

    - Avatar (2009),
    - Titanic (1997),
    - Jurassic World (2015),
    - Furious 7 (2015),
    - Frozen (2013).

3) Scatterplot matrices filtered by most and least expensive 1000No. movies were used to quickly review and visualise the relationships between different features in the dataset. The follow observations were made when comparing scaled budget, scaled revenue, scaled popularity, scaled runtime and vote average.

    - The positive correlation between budget and revenue suggests that there is a relationship between these factors, and that movies with higher budgets tend to have higher revenues. However, from the above analysis of profitable films, we know that outliers define a successful film or a box-office flop. The relationship (slope of correlation) whilst positive is not strong.
    - The positive correlation between budget and popularity suggests that there is a relationship between these factors, but it is less pronounced than the relationship between budget and revenue.
    - There is little correlation between budget/revenue and runtime, indicating that runtime is influenced by other factors.
    - There is little correlation between budget and vote average, indicating that other factors are at play in determining vote scores.

- The positive correlation between revenue and popularity suggests that there is a relationship between these factors, and that movies with higher revenues tend to have higher popularity. Again, the relationship is less pronounced.
- Overall, the analysis suggests that there are several factors at play in determining the success of a movie, including budget, revenue, popularity, runtime, and vote average, and that the relationships between these factors are complex and not fully understood.

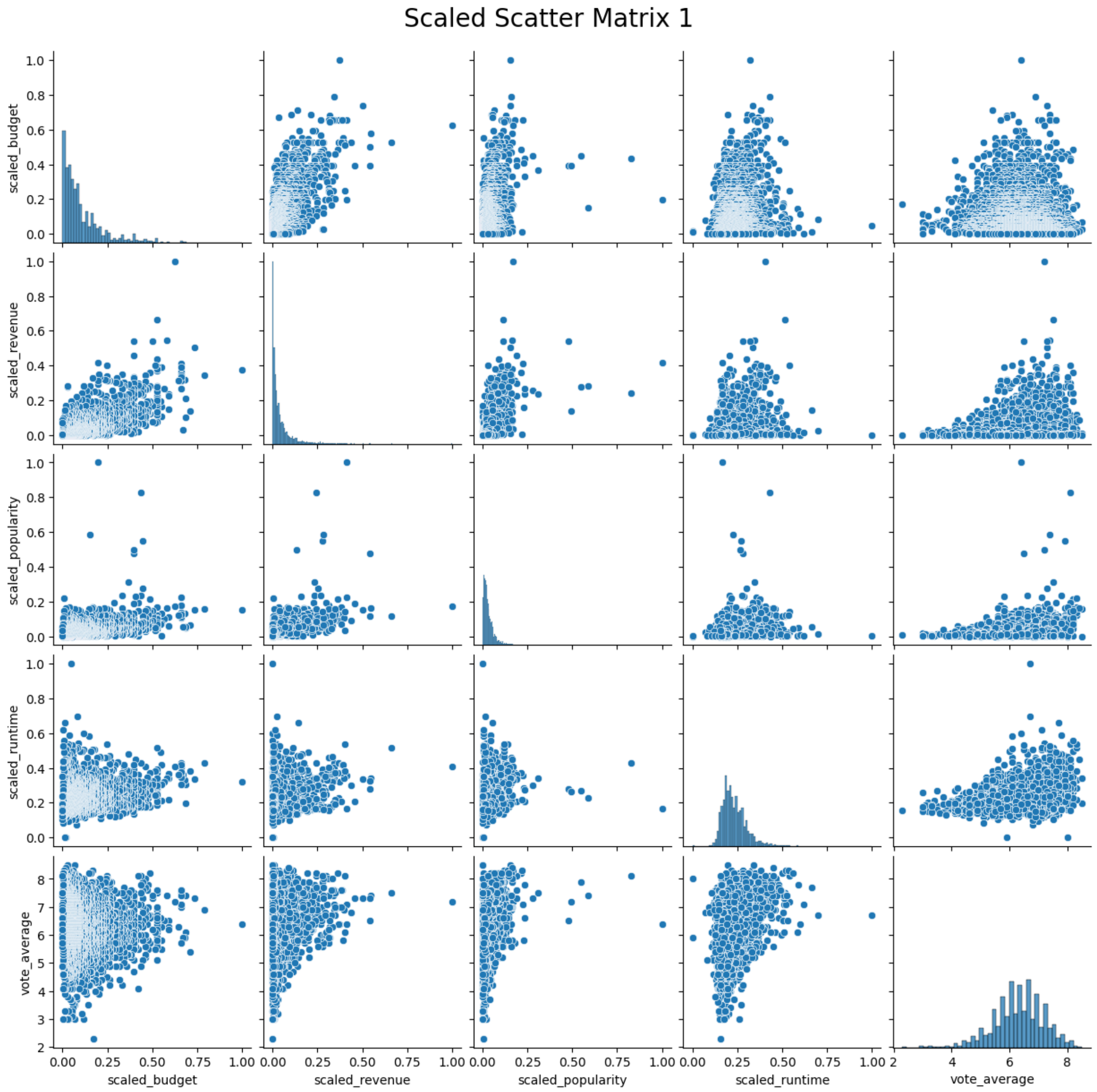


Figure 4: Scatterplot Matrix Scaled Features

4) Timeseries scatterplots were used to review the relationships between release year and different features in the dataset. The follow observations were made when

comparing release year, budget, revenue, popularity, runtime, and vote average. Scatterplots were selected over lineplots where appropriate to show outliers.

- The positive correlation between budget/revenue and year suggests that over time, film production costs and revenues have increased. However, this does not account for inflation and changes in purchasing power.
- The positive correlation between popularity and year suggests that improvements in promotion strategies have led to an increase in popularity, especially in recent years.
- The correlation between runtime and year is less pronounced, indicating that runtime is influenced by other factors.
- The average vote per year is declining, which may suggest that voters are becoming less engaged or that movies are becoming less innovative.

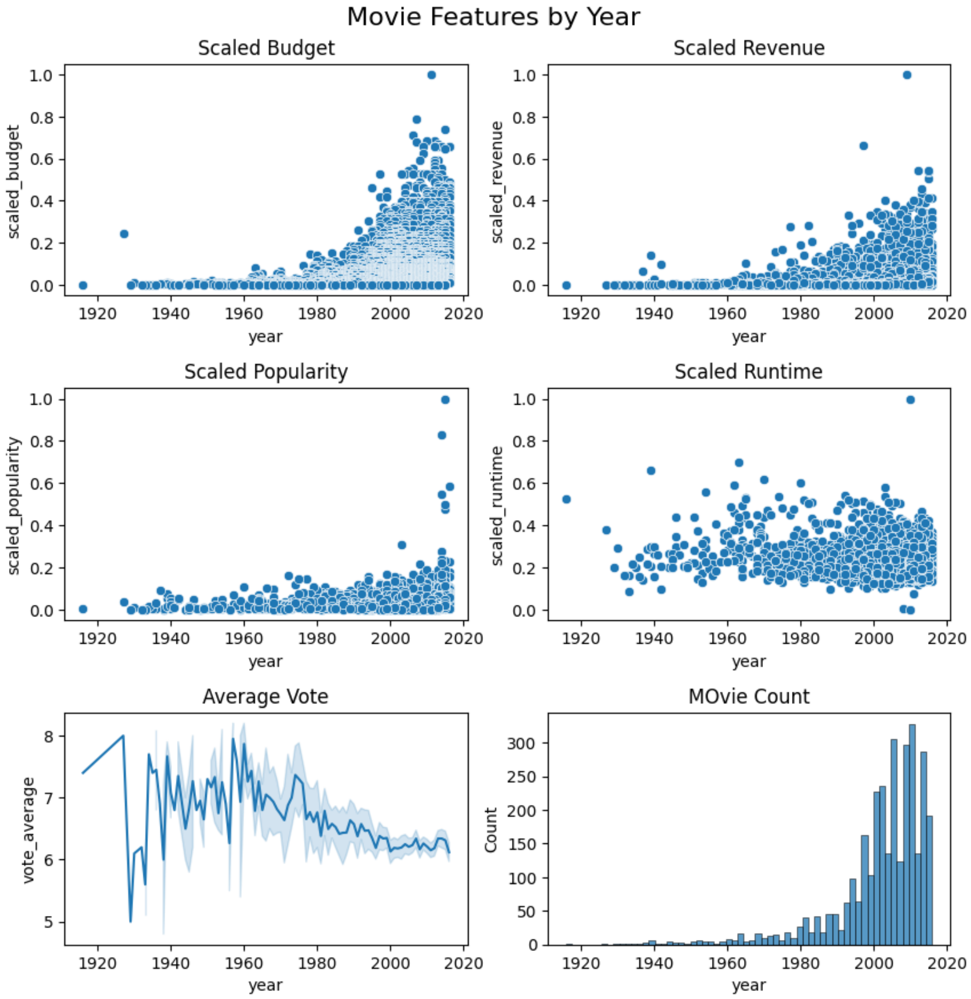


Figure 5: Scatter/Lineplots by Movie Feature and Release Year

5) Movies features per decade were used to spot for any dramatic changes or trends within the movie industry that may not be as obvious, especially those plots that were less pronounced.

- The movie industry experienced a large raise in budgets to revenue in the 1920s, with profits the lowest on record. This could be related to world economic events such as post-WW1 recovery and the Great Depression.
- The industry bounced back in the 1930s but dipped in the 1940s and 1950s due to factors such as the economic impacts of WWII and the rise of television.
- In the 1970s and 80s, budgets and revenues increased in tandem, but profits dipped. This could be due to changes in technology, film industry, or economic factors such as inflation or US economic policy. The scaled line plot shows a linear relationship between budget and revenue widen and become non-linear during this period. To today, the scaled budget and revenue relationship has not narrowed.
- Vote average mean has been declining per decade.

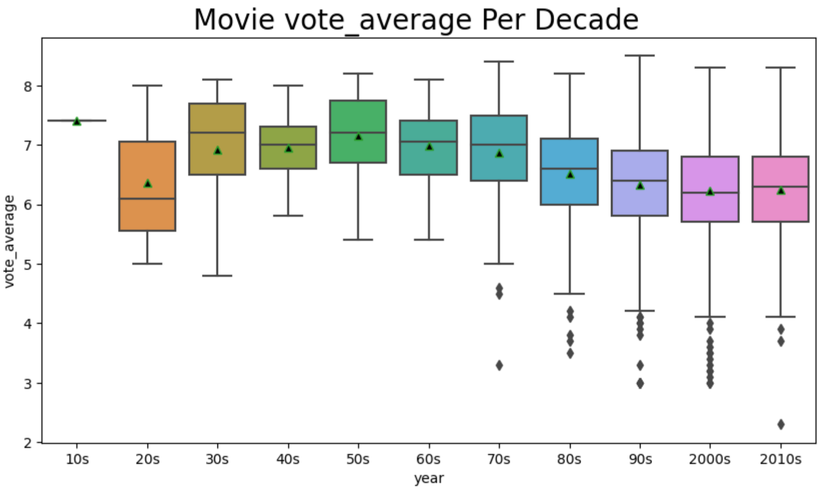


Figure 6: Boxplot of Movie Vote Average by Decade

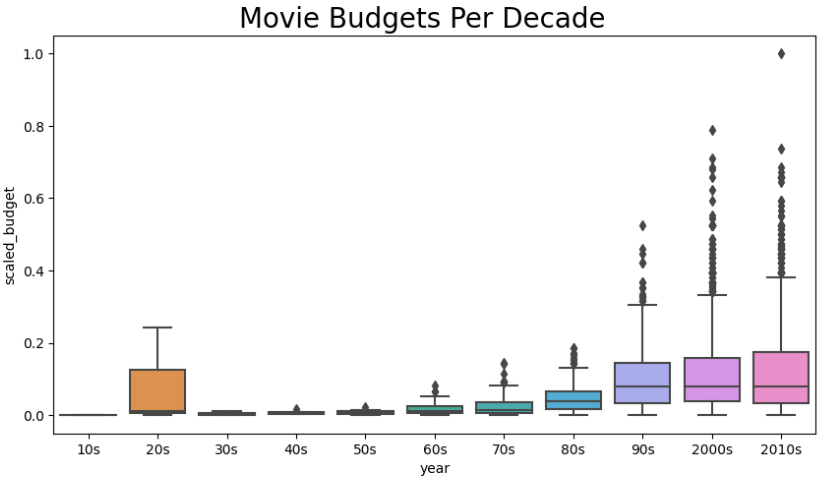


Figure 7: Boxplot of Movie Budgets by Decade

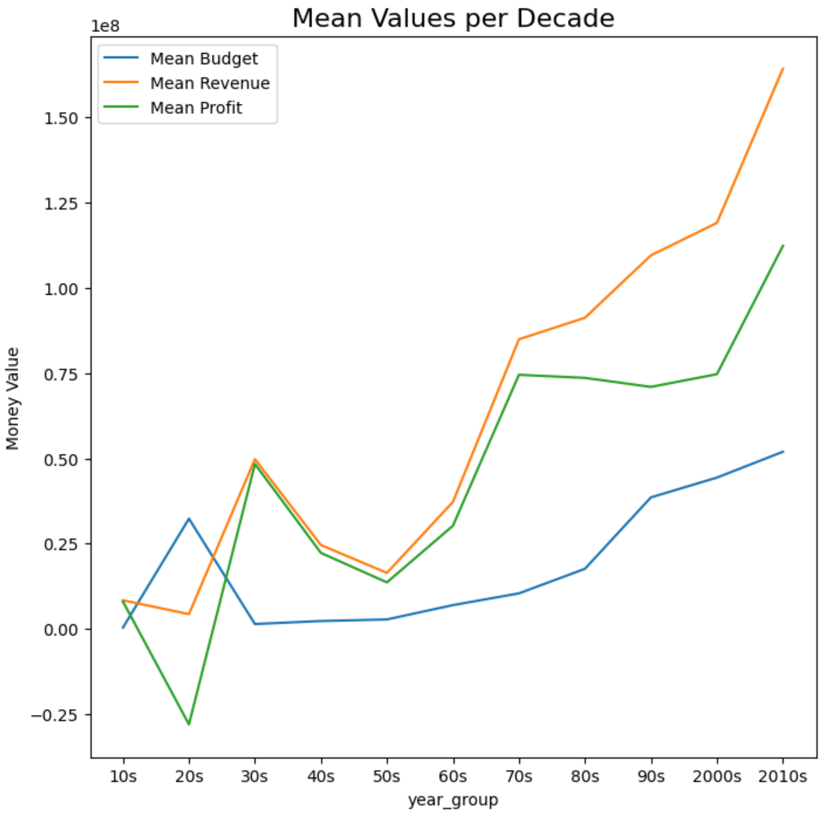


Figure 8: Lineplot of Budget/Revenue Means per Decade

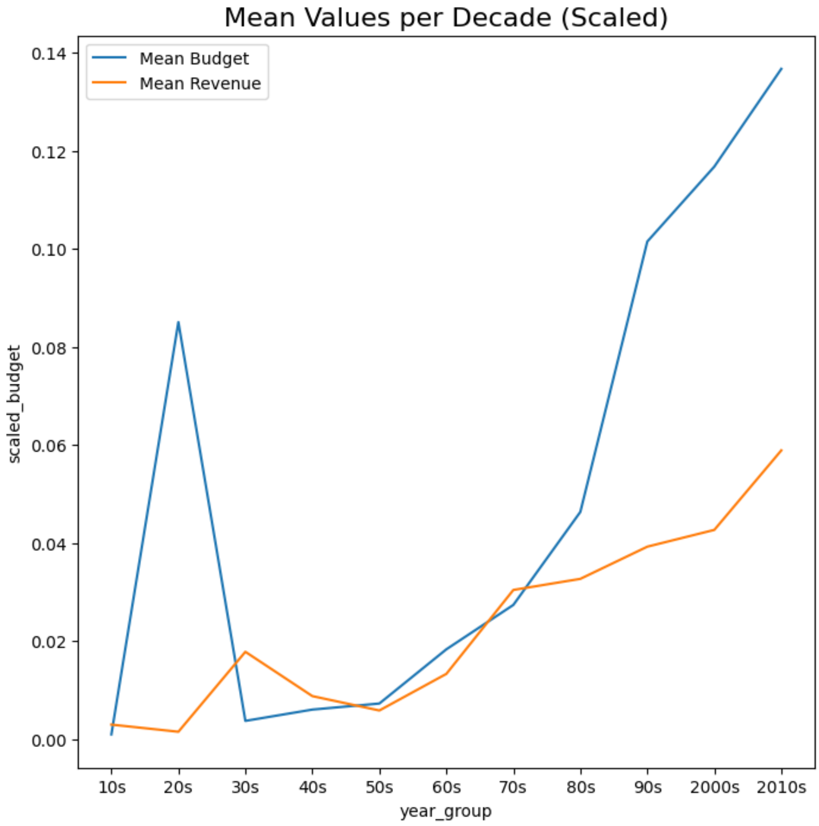


Figure 9: Lineplot of Scaled Budget/Revenue Means per Decade

6) What makes a film profitable or what is driving people to watch them is something that should be investigated. This will be investigated by reviewing genre data.

- The most populous genre in the dataset by frequency is Drama, followed by Comedy, Thriller then Action. In recent decades, these genres as the most populous.

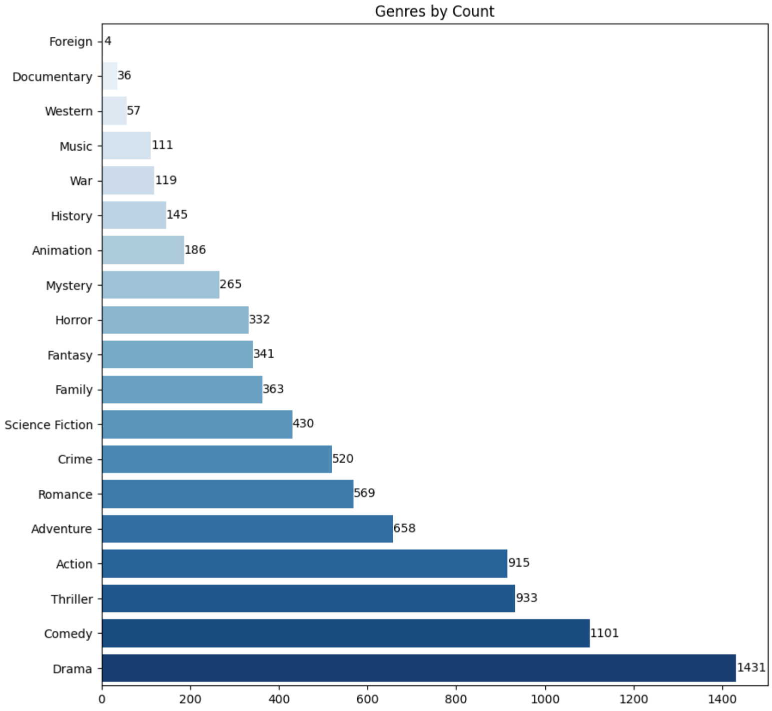
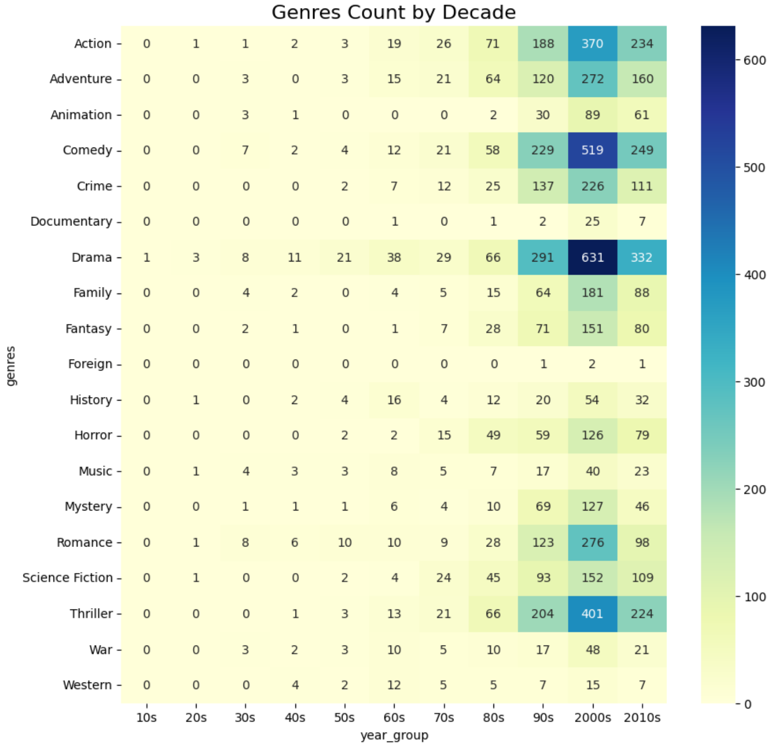


Figure 10: Barplot Genre Count



genres	10s	20s	30s	40s	50s	60s	70s	80s	90s	2000s	2010s
Action	0	1	1	2	3	19	26	71	188	370	234
Adventure	0	0	3	0	3	15	21	64	120	272	160
Animation	0	0	3	1	0	0	0	2	30	89	61
Comedy	0	0	7	2	4	12	21	58	229	519	249
Crime	0	0	0	0	2	7	12	25	137	226	111
Documentary	0	0	0	0	0	1	0	1	2	25	7
Drama	1	3	8	11	21	38	29	66	291	631	332
Family	0	0	4	2	0	4	5	15	64	181	88
Fantasy	0	0	2	1	0	1	7	28	71	151	80
Foreign	0	0	0	0	0	0	0	0	1	2	1
History	0	1	0	2	4	16	4	12	20	54	32
Horror	0	0	0	0	2	2	15	49	59	126	79
Music	0	1	4	3	3	8	5	7	17	40	23
Mystery	0	0	1	1	1	6	4	10	69	127	46
Romance	0	1	8	6	10	10	9	28	123	276	98
Science Fiction	0	1	0	0	2	4	24	45	93	152	109
Thriller	0	0	0	1	3	13	21	66	204	401	224
War	0	0	3	2	3	10	5	10	17	48	21
Western	0	0	0	4	2	12	5	5	7	15	7

Figure 11: Heatmap Movie Genre Production per Decade

- Those genres that have the highest mean budgets include Adventure, Animation and Fantasy. However, within the most recent decades of this dataset, Action and Science Fiction and Westerns have been increasing more expensive to make.

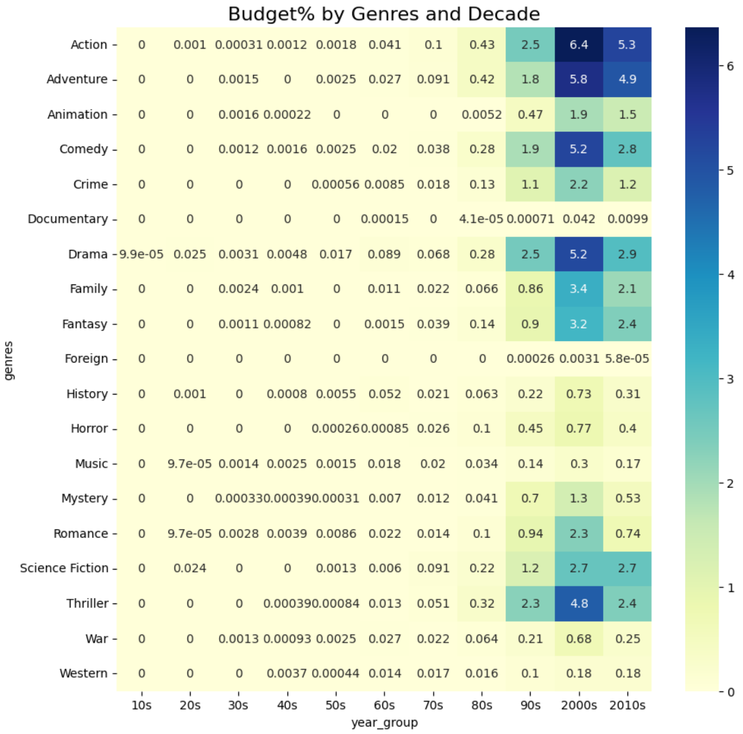


Figure 12: Heatmap Movie Genre by Budget%

- Those genres that the highest mean revenues include Animation, Adventure and Fantasy. However, within the most recent decades of this dataset Action, Drama and Thriller have been taking more revenue.

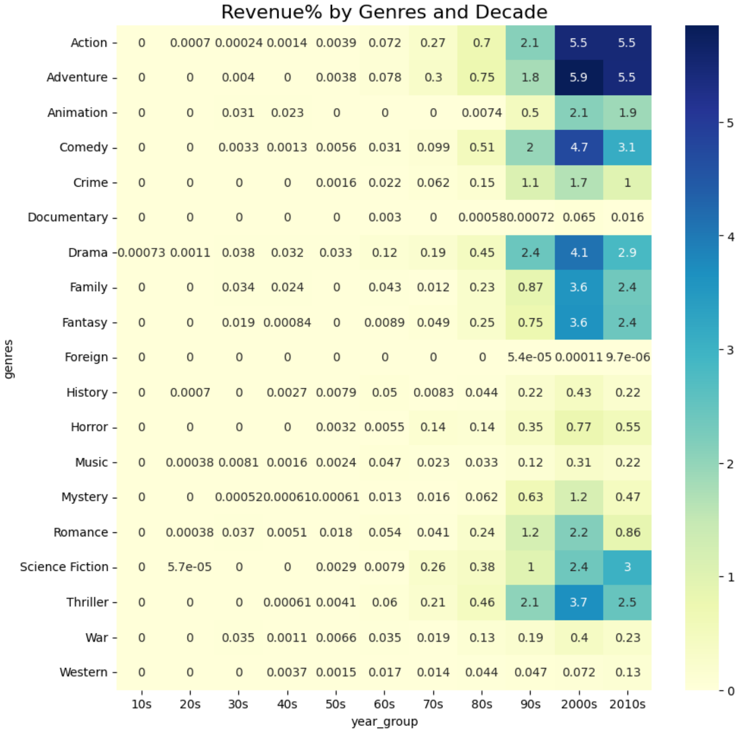


Figure 13: Heatmap Movie Genre by Revenue%

- Those genres that have made the largest profits include Adventure, Action, Fantasy, and Science Fiction. However, within the most recent decades of this dataset drama, comedy, and thriller have been making more profit.

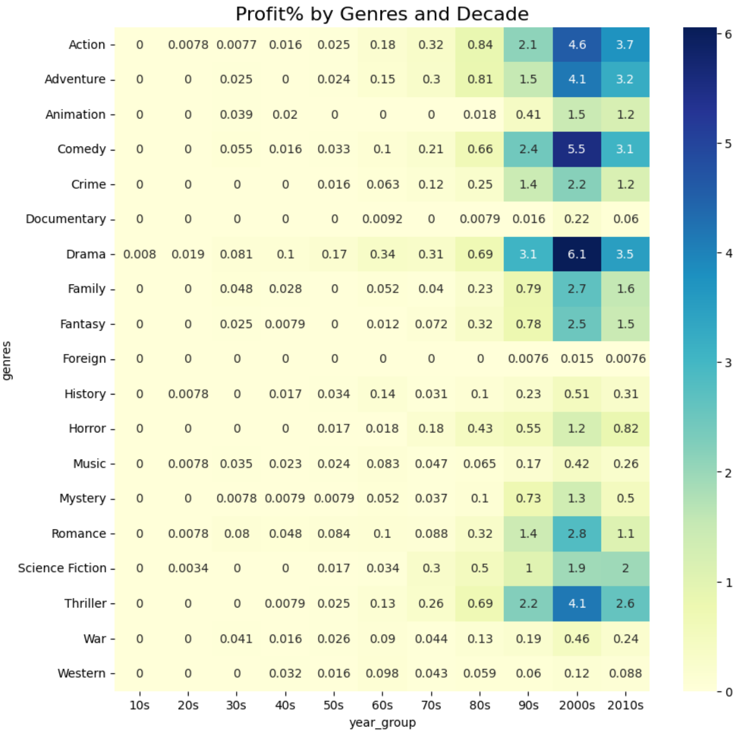


Figure 14: Heatmap Movie Genre by Profit%

- Those genres most popular by mean value include Adventure, Action, Fantasy, Science Fiction and Animation. However, within the most recent decades of this dataset Science-Fiction has been becoming more popular.

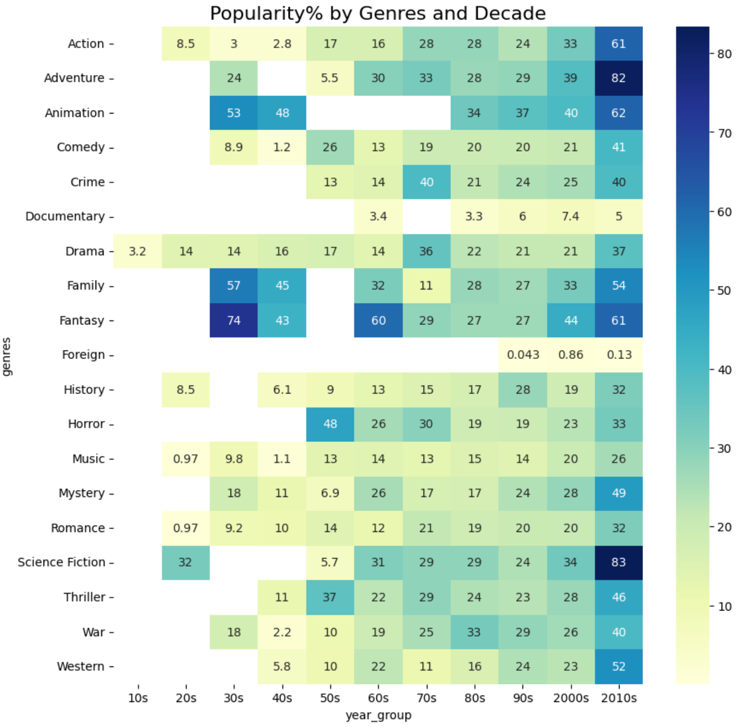


Figure 15: Heatmap Movie Genre by Popularity

- Those genres that have the highest mean vote average include War, History, Documentary, Western and Animation. It is notable that these have the lowest frequency counts but the highest mean vote average. However, within the most recent decades there is no stand-out genre, all declining in voter average. Foreign genre has increased slightly, and Horror has declining significantly.

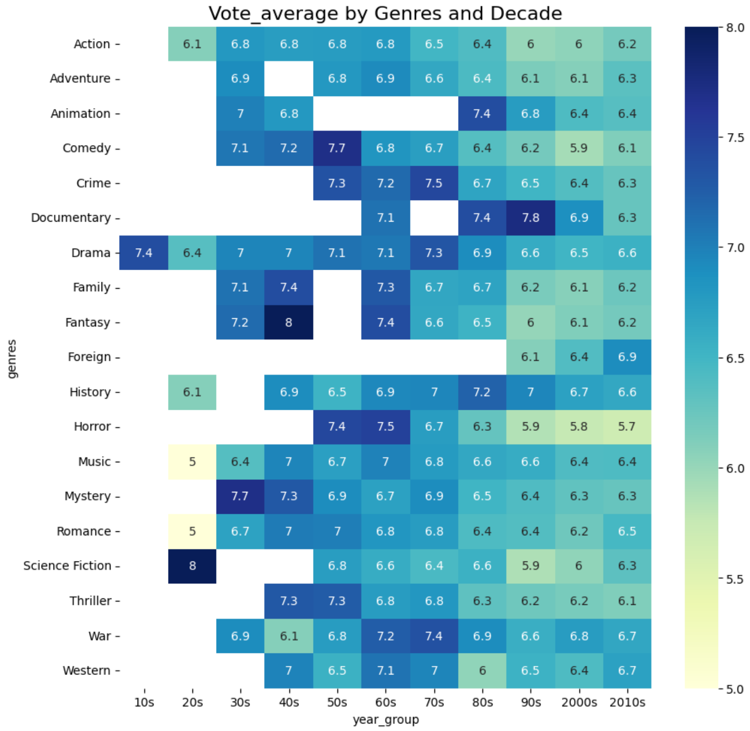


Figure 16: Heatmap Move Genre by Vote Average

Main Findings from Movies Dataset

1) Film Success – Budget vs. Revenue (Profitability).

In summary, the data shows that most movies have lower budgets, revenues, popularity, and runtime, with fewer movies having higher values. The vote average is normally distributed but with a longer left tail, indicating that voters tend to give slightly higher scores. The profitability of a movie is not solely determined by its budget and revenue, as there are other factors that play a role such as popularity and vote average. The scatterplot matrices reveal a positive correlation between budget and revenue, as well as budget and popularity. Overall, it is important to consider multiple factors when assessing the success of a movie, as we see above, known successful movies tend to be the outliers and do not typically follow the relationship trends indicated above. Other factors are responsible.

2) Movie Vote Average per Decade.

The data shows a downwards trend in vote average per decade for all movies with foreign genre movies slightly increasing and Horror declining significantly. It is unclear if this trend is due to an increase in the number of films being produced or if viewers are becoming more critical of movies. Further research is needed to determine the cause of this trend and what factors may be influencing it, such as changes in screenplay, actors, production, politics, and genre preferences.

3) Mean Values per Decade

The data shows that the movie industry experienced a significant increase in budgets to revenue in the 1920s, with profits at the lowest on record. This could be related to world economic events such as post-WW1 recovery and the Great Depression. The industry recovered in the 1930s but saw a decline in the 1940s and 1950s, due to factors such as the economic impacts of WWII and the rise of television. In the 1970s and 80s, budgets and revenues increased together but profits dipped. This could be due to changes in technology, the film industry, or economic factors such as inflation or US economic policy. The scaled line plot shows that the linear relationship between budget and revenue widened and became non-linear during this period and has not narrowed since then. More research is needed to fully understand this recent shift.

4) Vote Average per Genre

The most profitable film genres, such as Adventure, Fantasy, and Animation, do not always have the highest vote averages. War, History, and Documentary films, which tend to have lower revenues, often have higher vote averages. This could be due to these films dealing with real-life events and having a more targeted audience, leading to a higher quality of films being produced. However, it is important to note that War, History, and Documentary films have lower vote counts, which may influence these results. Further research is needed to understand the relationship between vote average and genre.

Recommendations for Future Analysis

1) Currency Standardisation

To understand profitability per movie feature better, next steps should include the standardising and adjusting for inflation. Current figures do not appear to consider such changes in purchasing power.

2) Other Features

Other features should be reviewed including for potential changes in technology, other media such as TV, famous actors/directors, and publicity budgets.

THIS REPORT WAS WRITTEN BY: Karl Gibson