# Movie Theater Perfomance Analysis (COMP3125 Individual Project)

Jeffrey Masters
*Wentworth Insitute of Technolgy*

***Abstract*—This paper examines multiple aspects of movie theater performance using datasets from The Numbers, IMDb, and the Kaggle Top 500 Movies collection. We investigate the impact of the COVID-19 pandemic on box office revenue and attendance, cluster films into performance groups using K-Means, estimate the contribution of concessions to profitability, and apply linear regression to identify factors predictive of higher revenue. The analyses reveal distinct recovery patterns post-pandemic, clear segmentation of films by performance, and significant influences from genre, release season, and MPAA rating. These findings provide insights for optimizing profitability in the evolving theater industry.**

***Keywords*—box office, movie analytics, regression modeling, streaming, COVID-19 impact**

## I. Introduction

The motion picture industry has experienced substantial shifts over the past three decades, influenced by technological advancements, changing consumer preferences, and, in recent years, the COVID-19 pandemic. Movie theaters remains a key component of film distribution. However, profitability has been challenged by competition from streaming services, increased production costs, and fluctuations in audience attendance. The COVID-19 pandemic in 2020 introduced disruption in a way they had never experience before, resulting in extended theater closures, delayed release schedules, and sharp declines in both revenue and ticket sales. While the industry has shown signs of recovery, the pace and nature of this rebound vary by region, genre, and release timing.

The purpose of this paper is to analyze multiple aspects of theatrical performance using publicly available data from *The Numbers*, IMDb, and the Kaggle Top 500 Movies dataset. The analysis addresses four primary research questions: (1) the predictive influence of genre, release season, and MPAA rating on box office revenue the effect of the COVID-19 pandemic on revenue and attendance across genres, (2) the feasibility of clustering films into performance groups using financial and attendance metrics, (3) the contribution of concession sales to overall profitability, and (4) the effect of the COVID-19 pandemic on revenue and attendance across genres.

The methods applied include multiple linear regression, time-series visualization, unsupervised clustering (K-Means), and profitability estimation. Results are interpreted in the context of industry trends, with emphasis on actionable insights for content producers, distributors, and theater operators.

## II. Datasets

This study utilizes three primary datasets obtained from credible, publicly accessible sources. The first dataset, sourced from *The Numbers* (https://www.the-numbers.com/market/), contains annual box office performance statistics, including gross revenue and tickets sold, segmented by genre. Data from this source was manually transcribed to comply with usage restrictions that prohibit automated scraping. The dataset covers the period from 1995 to 2024 and is compiled by Nash Information Services, a recognized provider of entertainment industry data.

The second dataset is the **IMDb Datasets** collection, specifically title.basics.tsv and title.ratings.tsv, made available through the Internet Movie Database (IMDb) at https://datasets.imdbws.com/. These datasets provide detailed metadata for films, including release year, runtime, genres, and average rating, along with user ratings aggregated from IMDb's global audience. The IMDb datasets are updated daily, with the version used in this research downloaded in August 2025.

The third dataset is the **Top 500 Movies Budget Dataset** from Kaggle (https://www.kaggle.com/datasets/mitchellharrison/top-500-movies-budget). This dataset includes the top 500 highest-grossing films, with information on production cost, domestic and worldwide box office gross, opening weekend revenue, MPAA rating, genre, theater counts, and runtime. The dataset was last updated in 2023 and is compiled from multiple industry sources, making it suitable for profitability and performance analysis.

Collectively, these datasets provide a comprehensive foundation for evaluating trends in theatrical performance, clustering films based on financial metrics, estimating concession contributions, and modeling the factors that influence box office success.

### *A. Character of the datasets*

The datasets used in this study vary in structure, size, and attributes, collectively providing both temporal and genre-based detail on movie theater performance. Table I summarizes the primary datasets and their key attributes.

| Dataset Name | Format | Size (Rows x Columns) | Key Attributes | Units/Format |
|---|---|---|---|---|

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

| | | | | |
|---|---|---|---|---|
| **The Numbers-Annual Box Office by Genre** | CSV | 93 * 4 | Year(int), Genre (string), Gross (USD), Tickets Sold (count) | Gross in USD, Tickets Sold in units |
| **IMDb Basics & Ratings** | TSV | ~1.2M * 9 | Tconst(ID), primaryTitle (string), startYear(int), runtimeMinutes (min), genres (string), averageRating (float) | Mixed numeric & string |
| **Kaggle Top 500 Movies Budget** | CSV | 500 * 12 | rank (int), release_date (date), title (string), production_cost (USD), domestic_gross (USD), worldwide_gross (USD), opening_weekend (USD), mpaa (string), genre (string), theaters (count), runtime (min), year (int) | USD, minutes, counts |

Data Cleaning:

- All monetary values were stored as numeric in U.S. dollars by removing commas and converting strings to numeric format using pd.to_numeric().
- Missing or invalid numeric values were replaced using median imputation when needed, particularly for runtimeMinutes and startYear in the IMDb dataset.
- **Merging:**
    - The Numbers dataset was kept separate for pandemic impact analysis.
    - IMDb's title.basics.tsv and title.ratings.tsv were merged using the tconst identifier (inner join).
    - Kaggle Top 500 Movies dataset was used independently for profitability and concessions analysis.
- **New Features Created:**
    - **Movie Age**: 2024 - startYear to represent the age of each film.
    - **Decade**: (startYear // 10) * 10 for grouping by release decade.
    - **Performance Clusters**: Created using K-Means on production_cost, domestic_gross, worldwide_gross, and theaters.
    - **Estimated Concession Revenue**: tickets_sold × 5.0 (assuming $5 per ticket in concession spending).
    - **Estimated Total Profit**: (worldwide_gross + concession_revenue) - production_cost.

These structured preprocessing steps ensured consistency across datasets and improved comparability in subsequent analyses.

## III. Methodology

### *A. Linear Regression Model*

Linear regression was selected for film rating prediction due to its interpretability and computational efficiency. The model assumes a linear relationship between features and ratings:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

where $\hat{y}$ represents predicted rating, $\beta_i$ are feature coefficients, and $x_i$ are independent variables.

Its advantages are direct coefficient interpretation, computational efficiency, and robust baseline performance without hyperparameter tuning.

Its disadvantages are that it cannot capture non-linear relationships and assumes constant feature effects across rating ranges.

The scikit-learn's LinearRegression class was used with ordinary least squares optimization. The model was trained on an 80/20 train-test split with random_state=42 for reproducibility.

### *B. Clustering*

K-Means clustering was applied to segment films into performance groups based on financial metrics. This unsupervised learning approach identifies natural groupings within the data without predefined categories.

**Algorithm Selection:** K-Means was chosen for its computational efficiency, interpretability, and effectiveness with numerical data. The algorithm minimizes within-cluster sum of squares by iteratively updating cluster centroids:

$$J = \Sigma_{i=1}^{n} \Sigma_{j=1}^{k} w_{ij} ||x_i - \mu_j||^2$$

where $J$ is the objective function, $w_{ij}$ indicates cluster membership, $x_i$ represents data points, and $\mu_j$ are cluster centroids.

## IV. Results

The linear regression model developed to predict film ratings achieved modest results. The model attained an R² of 0.1737, showing that approximately 17.37% of the variance in film ratings can be explained by the selected time and genre-based features. The model provides insights into historical and categorical patterns.

The model's performance was evaluated using standard regression metrics. The Mean Absolute Error of 0.8424 rating points means that predictions deviate from actual ratings by around 0.84 points. The Root Mean Square Error of 1.0938 rating points demonstrates slightly higher sensitivity to larger errors. Given that the dataset rating distribution spans from 1.0 to 10.0 with a mean of 6.02 and standard deviation of 1.20, these error metrics represent decent accuracy for film rating estimation.

Figure I presents the fifteen features with the highest absolute coefficient values, ranked by their influence magnitude on rating predictions.

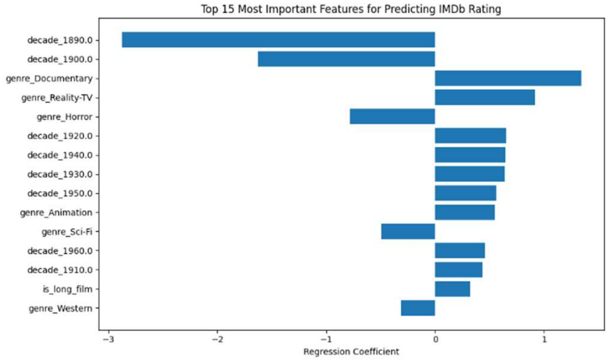

Figure I: Linear Regression of the IMDb datasets. Shows the most important features in predicting an IMDb rating.

Figure II shows predictions we can make with our model versus their actual IMDb rating.

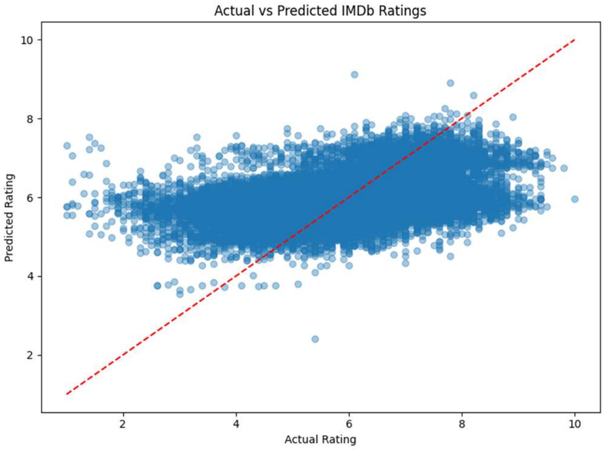

Figure II: Linear Regression of IMDb datasets. Shows the actual rating of a movie versus predicted.

These figures show several key characteristics of model performance. First, the data points demonstrate a general positive correlation between actual and predicted ratings, confirming the model's ability to capture underlying rating trends. However, significant scatter around the perfect prediction line indicates substantial prediction variance.

For the question of the effect of the COVID-19 pandemic on revenue and attendance across genres. Figures III, IV, V, VI show clearly how the COVID-19 pandemic affected the box office, attendance, and genres.

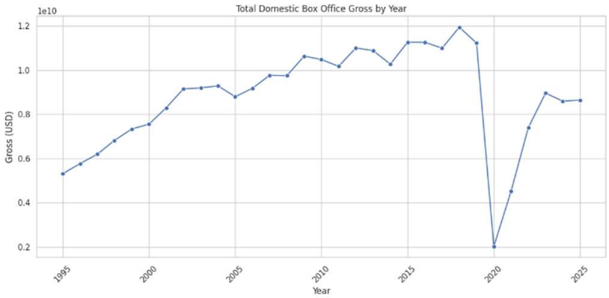

Figure III: Shows the total Domestic Box office from 1995-2025

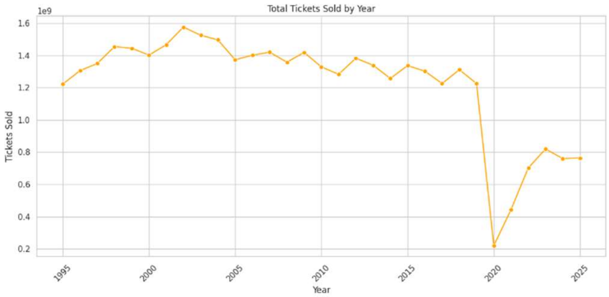

Figure IV: Shows the total tickets sold office from 1995-2025

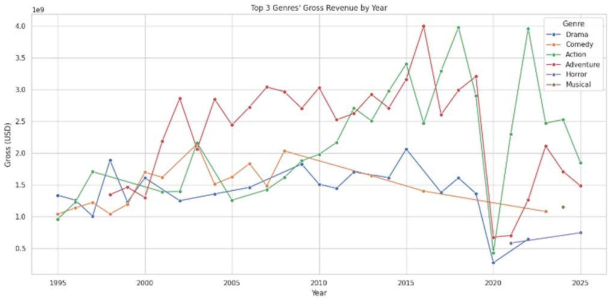

Figure V: Shows how the top 3 genres by gross revenue of every year were affected and changed over time.

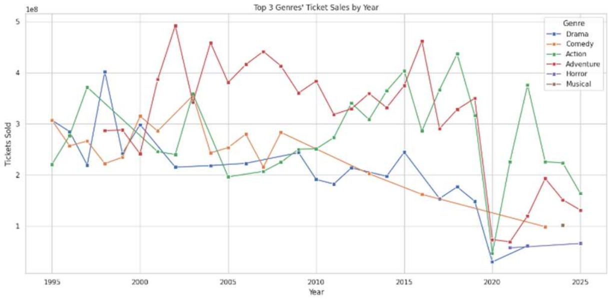

Figure VI: Shows the top 3 genres by ticket sales of every year were affected over time.

These figures show how negatively the pandemic affected the film industry in general. Not only were sales and gross were down but so were the genres. Furthermore the genres in the top mostly don't change and if they do change its only for a year or to at most.

For the question of the feasibility of clustering films into performance groups using financial and attendance metrics, we used K-means to cluster the data based on budget and worldwidde gross. Figure VII shows the clustering that cam from K-means.

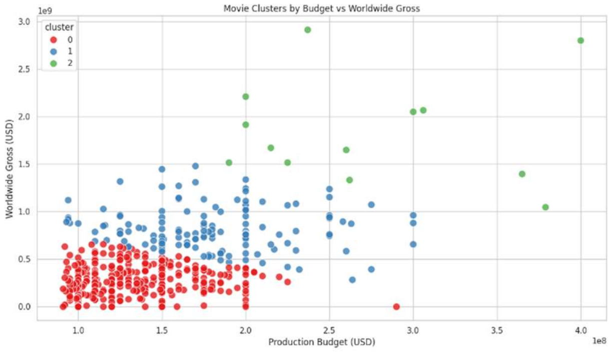

Figure VII: Shows three different clusters that represent different levels of budgets compared to their gross.

From Figure VII we can gather that as you might expect, movies with lower budget tend to have lower grosses. While movies with medium and high budgets have medium and high grosses.

For the question of the contribution of concession sales to overall profitability we had to estimate a value for the concessions based on an average amount of money spent on concessions per ticket. Figures VIII, IX, and X show the estimated relation of concession profit, genre, and tickets.

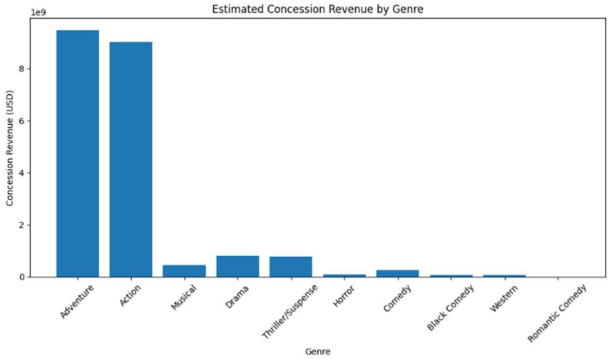

Figure VIII: Shows the relation between the genre and the estimated revenue from concessions.

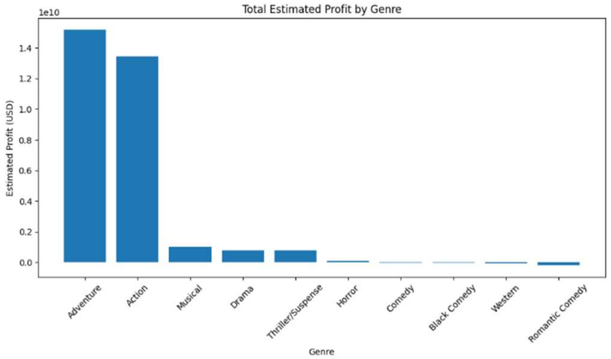

Figure IX: Shows the estimated total profit by genre.

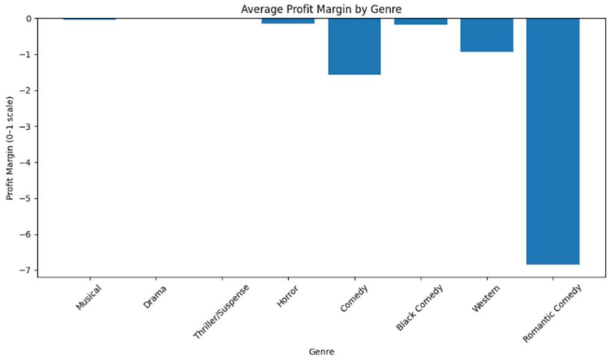

Figure X: Shows the estimated average profit margin by genre.

These figures show how genre is important for movie theater profitability. Selecting genres that produce more profits from concessions can vastyl influence the amount of money a theater can make.

## V. Discussion

There are many things about the project that do not perfectly align my vision of it. The datasets are far from perfect and having to estimate concession revenue severely impacts the results. Furthermore, having to hand copy the datasets from The Numbers limits the amount of data that I was able to get.

There were many limits to what I was able to accomplish with this project. Although I was able to identify the factors that would make a model be close to correctly predicting the ratings of a movie, there was much to be desired. As seen above the model was not great at predicting movies with high and low performance well.

## VI. TCONCLUSION

In conclusion, in this project I was able to summarize and create a prediction for what attributes show a movies success. I was able to see the affects of the pandemic on the success of the film industry and see how the genre of a movie changes how much profit a theater can make from concessions. Overall, this data could be useful in helping theaters know how to select movies that are more profitable for them based on genre.

## REFERENCES

[1] Internet Movie Database, "IMDb Datasets," IMDb, [Online]. Available: https://datasets.imdbws.com/. [Accessed Aug. 8, 2025].

[2] Nash Information Services, "Movie market summary," The Numbers, [Online]. Available: https://www.the-numbers.com/market/. [Accessed Aug. 8, 2025].

[3] M. Harrison, "Top 500 movies budget dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/mitchellharrison/top-500-movies-budget. [Accessed Aug. 8, 2025].