Dataset Analysis

By: Respeto Sofia Amihan Molase

Content Page

1.0 Introduction

The Ultimate Film Statistics Dataset
A comprehensive movie statistics dataset compiled from multiple sources such as Wikipedia, The Numbers, and IMDB. This dataset will be used to conduct a box office analysis to analyse the relationships between production budgets, domestic and worldwide gross earnings, and profitability. There is a total of 4380 rows and 14 columns. Due to the complexity, depth, and imbalance of this dataset, I will be utilising tree-based machine learning algorithms that are more suited to these, datasets where unpredictability and outliers are key features of the theatre statistics scene. By the end of the project, a model that can predict the Worldwide Gross Revenue($) will be produced.

Features	**Datatype**	**Impression**
movie_title	Nominal	Nil
production_date	Ordinal	It needs to be separated into 3 individual columns to identify hidden patterns. Even though these would be split into discrete values, due to the large number of, for example, years, it is appropriate to use a histogram to chart the progression over time.
genres	Nominal	Genres can heavily influence the results of a movie due to differing popularities.
runtime_minutes	Continuous	Nil
director_name	Nominal	Nil
director_professions	Nominal	Nil
director_birthYear	Discrete	Nil
director_deathYear	Discrete	Contains mixed data types.
movie_averageRating	Continuous	Nil
movie_numerOfVotes	Discrete	Nil
approval_Index	Ordinal	A normalised indicator (on a scale of 0-10) is calculated by multiplying the logarithm of the number of votes by the average user rating. It provides a concise measure of a movie's overall popularity and approval among online viewers, penalising films that got too few reviews and blockbusters that got too many.
Production budget $	Continuous	Nil
Domestic gross $	Continuous	Nil

Worldwide gross $	Continuous	Nil

As an amateur film enthusiast, this project will allow me to gain valuable insights into the film industry by approaching it from a more analytical and data-driven perspective. This project will give me the opportunity to enhance my understanding of the complexities involved in machine learning model development, particularly in grasping the significant impacts of seemingly minor details such as parameter tuning. I hope to gain a deeper appreciation for the film industry, recognising the inherent challenges and unpredictability of box office performance and the presence of outliers.

Assumptions:

- A film’s revenue can be estimated using features such as budget, genre, runtime, production date, approval index, and the director involved
- Data from 2020 onwards may be affected by the COVID-19 pandemic, potentially introducing anomalies due to lockdowns, theatre closures, and shifts in consumer viewing behaviour toward streaming platforms

Hypothesis:

- Domestic gross revenue will have the strongest correlation with the target variable, worldwide gross revenue, as strong local performance usually translates to global success
- Action and adventure films are more likely to generate higher revenue compared to other genres, as they tend to appeal to a global audience and benefit from larger marketing budgets
- Films with moderate and reasonable runtimes are expected to perform better, as they strike a balance between audience engagement and accessibility, leading to higher viewership
- Films released during holiday seasons, particularly in December, are expected to achieve the highest box office performance due to increased leisure time and holiday-themed marketing strategies

2.0 Data Exploration and Pre-Processing of Data

2.1 Data Discovery and Profiling

The process of gathering and analysing data sources to assess their **relevance**, explore **initial patterns**, understand their **structure**, as well as hypothesise potential **relationships**. This includes a thorough evaluation of data quality, focusing on **completeness**, **consistency**, and **distribution**. During this process, the identification of outliers, missing values, or any other potential issues can be conducted to ensure the integrity and reliability of the dataset before curating the relevant machine learning model.

1. **Final Testing Data Extraction**
   Two untouched and unseen rows are extracted from the bottom of the original dataset for final testing on the website.
2. **Feature Standardisation**
   Reviewing and converting the feature headings into a standardised snake casing format for consistency and readability.
3. **Dataset Preview**
   A preview of the first 5 rows, the last 5 rows, and 5 randomly selected rows to gain an initial understanding of the dataset. This allows the formation of assumptions, identification of the general structure, as well as the verification of data quality. Including the last 5 rows provides for an assessment of consistency throughout the dataset, while random sampling helps to spot-check potential errors that may not be visible in sequential rows, offering a more representative view.
4. **Dataset Dimensions**
   To obtain an understanding of the scale of the data and to assess its computational feasibility for hyperparameter tuning.
5. **Data Type Examination**
   Differentiation of qualitative and quantitative features is needed to determine the appropriate visualisation techniques for future exploratory data analysis.
6. **Structural Analysis**
   Checks for null values, nan values, missing values, infinite values, and unusual data types that might impact the quality of data.
7. **Missing Data Representation**
   Instead of leaving fields blank, the dataset utilises placeholders such as ‘-’ and ‘\N’ to indicate missing data. Identification of these rows are needed to assess the percentage of completeness.
8. **Data Aggregation**
   An overview aggregation of the relevant qualitative and quantitative values are needed to spot patterns and trends.
9. **Data Consistency**
   Checks for inconsistent inputs or duplicates within the dataset.
10. **Unique Value Analysis**
    Ensures that data aligns with the expected patterns and distributions.
11. **Statistical Summary**

Metrics such as count, mean, standard deviation, minimum, maximum, median, and quartile ranges are provided for statistical analysis. A non-standardised version is also created to display the raw values that make it easier to spot the deviations.

12. **Correlation Matrix**
    Insights into potential dependencies and influential factors is given through the overview of relationships between numerical variables.
13. **Interesting Observations**
    a. 3 different movies happen to have the same title, production date, production, and revenue despite being completely unrelated.
    b. A number of movies happen to have the same title yet produce different results
    c. **Outlier**: An abnormally long movie. (Solved through Discretisation-Binning Below)
    d. The movie with the highest budget does not happen to be the movie with the highest worldwide gross
    e. If the movie with the highest budget is not the most successful, then which movie is?
14. **Display Genre Outlier**
    Despite the dataset having a multitude of numerical outliers, most of these numbers are exceptions or have reasonable explanations as to why we still need to include them with alternative methods in dealing with them being more appropriate. However, having an outlier for the genre, a column where we would have to perform one-hot encoding eventually, would be risky to include as it can skew the results greatly.

Action	Comments
Final Testing Data Extraction	Nil
Feature Standardisation	Nil
Dataset Preview	- Field formatting has to be done for production_date and genres, as well as the rows that have filler blanks such as ‘-’ and ‘\N’ - Since we’re predicting worldwide gross, personal information related to the director is irrelevant based on domain knowledge - Feature extraction is derived from ratings and votes to curate the approval index hence, these two are redundant columns - Worldwide gross revenue consists of domestic gross revenue, showing signs of data leakage
Dataset Dimensions	(4378, 14)
Data Type Examination	- Categorical data is present and may need one-hot encoding
Structural Analysis	- No missing data or null values that require immediate attention - All the data types are as expected
Missing Data Representation	- The columns that include these filler blanks are

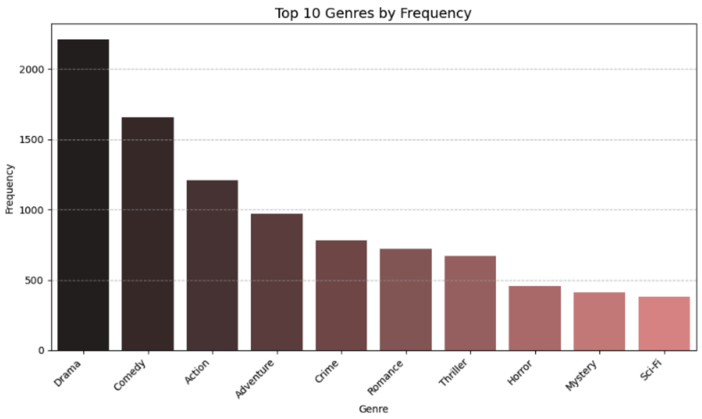
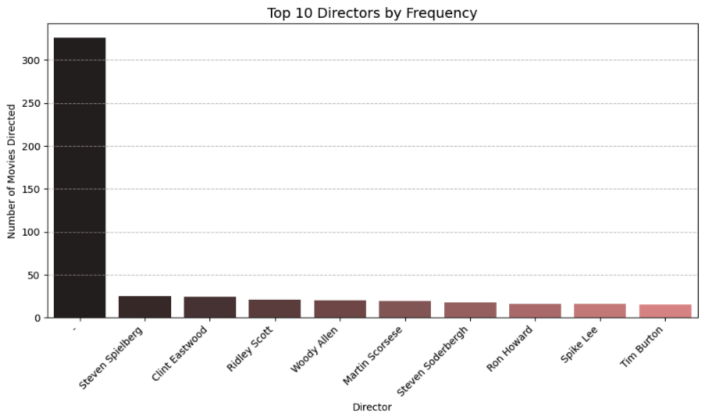
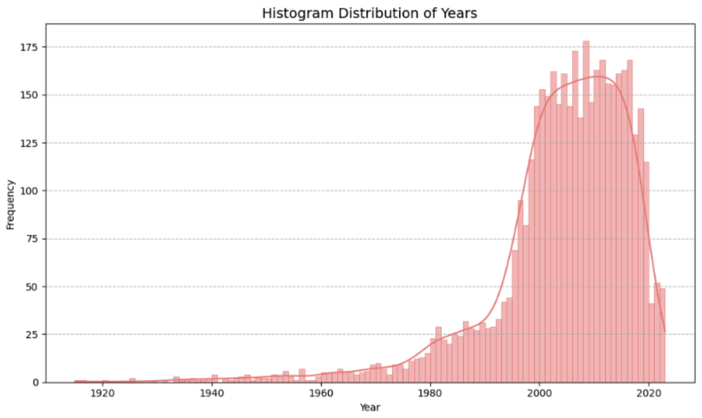
	irrelevant to our model building hence we can remove these columns instead of needing to fix these fields
Data Aggregation	- Even though successful movies are also related to the popularity of the director, there are too many movies that have their director as unlisted hence we may need to remove the column due to this imbalance - Votes seem to have a large range but since it is a redundant column, we need not worry
Data Consistency	- Checking that the frequency given on the rating has enough variation, which will affect how appropriate approval index may be
Unique Value Analysis	- There seems to be too many unique values for genres: 354
Statistical Summary	Runtime_minutes - There is an abnormally long movie of more than 4hours - The percentiles indicate that the majority of movies are between 96 and 120 minutes long - Due to the above distribution, binning the movie lengths into short, medium (for the majority), long, and very long (for the outliers) would be the most appropriate action Votes - The high standard deviation indicates that while most movies have a moderate number of votes, some extremely popular movies get a much larger share of the votes, skewing the distribution. This can also be seen from the range of votes. Budget - The distribution suggests that while most movies have a budget between $10 million and $50 million, there are a few blockbuster movies with very high budgets (e.g., $460 million). - Movies that are generally higher in budget do better due to the quality of the film. Although they are rather rare cases, it is necessary to include a higher budget, as it is more likely to indicate a higher worldwide gross, thus aligning with the relationship. (Proven below) Worldwide gross - As with domestic gross, the wide range and high standard deviation reflect a few movies that have earned substantially more than others, contributing to the overall average - Likewise, there seem to be a few movies that did not perform as expected due to the low minimum. Overall, the outliers in domestic and worldwide grosses, as well as in budget, suggest the presence of a few extremely successful and expensive movies that skew the averages.

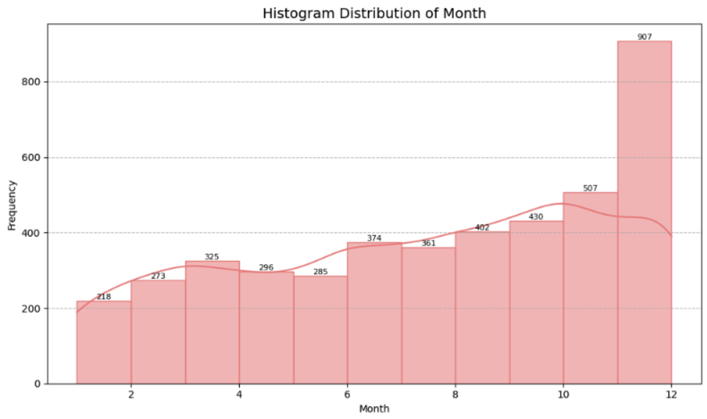
Correlation Matrix	- Since the approval index is derived from ratings and votes, there is a strong correlation present - A moderate correlation of votes and worldwide gross of 0.58, which suggests that movies with high votes tend to earn more - Budget and worldwide gross have a strong correlation of 0.73, which proves my previous statement - Domestic gross and worldwide gross have a correlation of 0.94, which may indicate data leakage Unexpected: - Weak correlation between budget and rating of 0.07 - Weak correlation between approval index and budget of 0.28
Interesting Observations	More information on Jupyter Notebook
Genre Outlier	After exploring the dataset and multiple data cleaning iterations, I noticed that only one movie had a genre listing of 'News', and therefore acts as an outlier where the row cannot be used for training or testing due to its singularity. As such, we would have to remove this row during data cleaning.

2.2 Exploratory Data Analysis (EDA)

Visualisation and analysis of the data is conducted to identify key **relationships**, **patterns**, and **trends**. These visualisations help highlight **correlations** and **outliers** that can be dealt with during cleaning and pre-processing. Each method of visualisation has been made in accordance with all the features and their datatypes, which were previously identified.

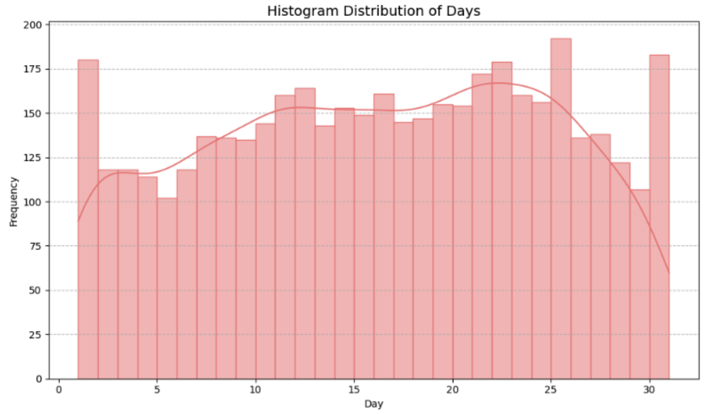
1. Individual Frequency

	**Top 10 Genres** From the bar chart, we can see that the data may be **imbalanced** as there is an uneven number of representatives for each genre. The most commonly occurring one would be drama.
	**Top 10 Directors** From the bar chart, we can see that the data is **highly imbalanced**, with many movies not having a director associated. This may affect our model building process and it would be best not to include this feature. Even though directors such as Steven Spielberg may produce higher grossing movies, there are other, more reliable features we can use in our model.
	**Years** From the histogram, we can see that as the years go on and the film industry evolves, there have been more and more films being produced. This could indicate that the **more recent the movie**, the **higher the quality** due to technological advancements, and thus the potential **increase in worldwide gross**. There is also a noticeable drop around 2020, most likely due to COVID-19.



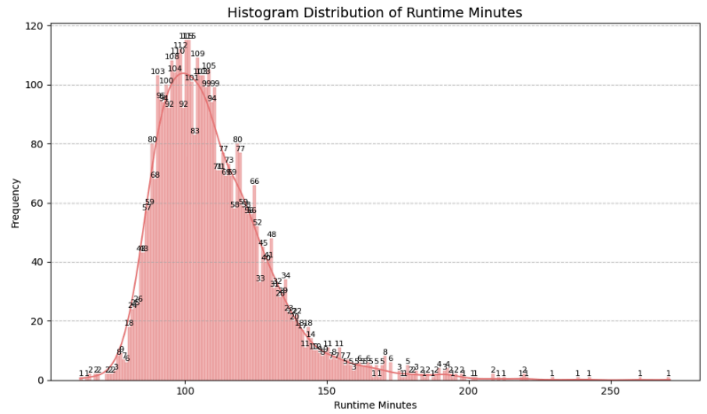
Months

From the histogram, we can see that during the holiday seasons, particularly December, more movies are being released.



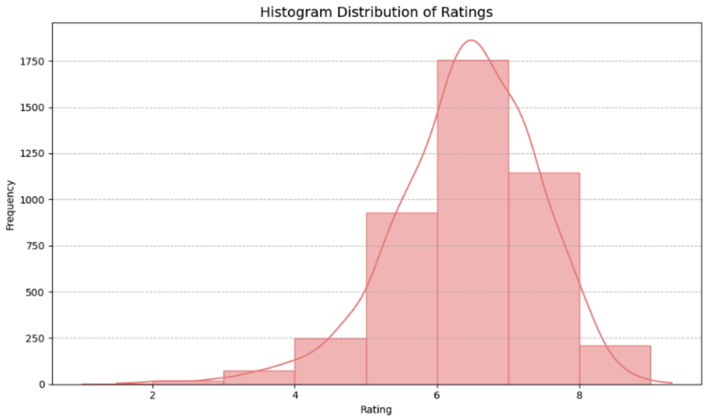
Days

From the histogram, we can see that the days take on a **relatively normal distribution**. It is unlikely that this feature will be impactful in the prediction of Worldwide Gross $.



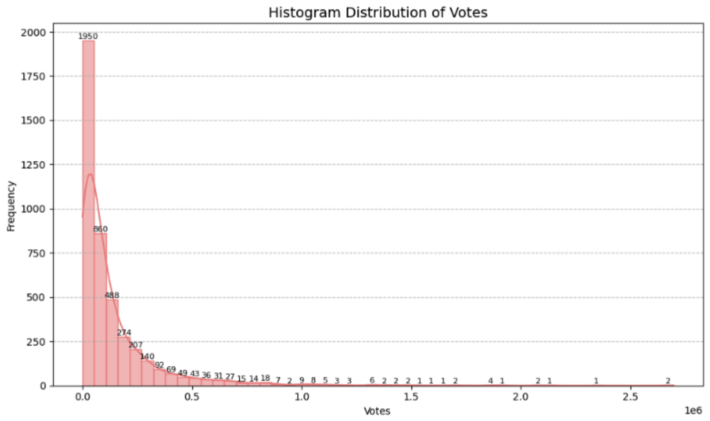
Runtime Minutes

From the histogram, we can see that it is **skewed to the left**, with the dataset containing a number of outliers. Since these are actual movies instead of field errors, it may be more appropriate to scale the feature incase any hidden patterns arise instead of omitting these rows.



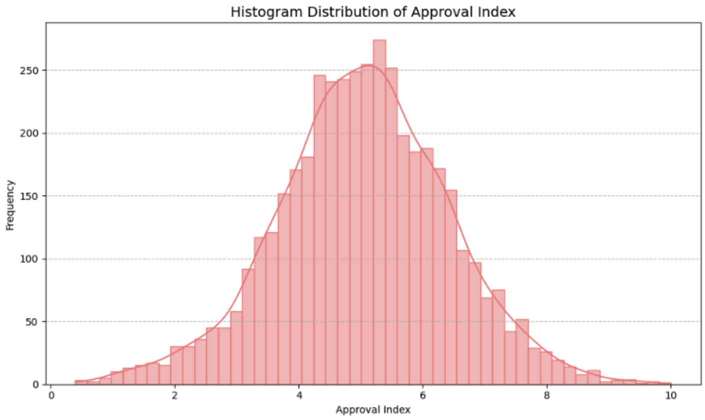
Ratings

From the histogram, the distribution for ratings is **rather normal**, maybe **slightly skewed to the right**. This could be due to the presence of some poorly rated movies, which are important in our model.



Votes

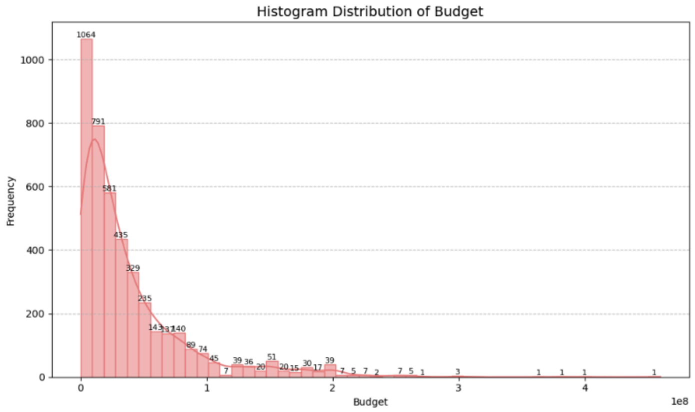
From the histogram, the distribution for votes is **highly skewed to the left** with a number of outliers.



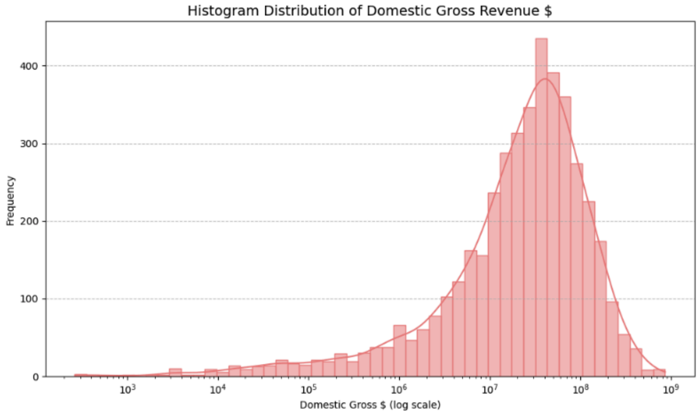
Approval Index

This feature is derived from the Ratings and Votes, taking on a **normal distribution**, therefore fixing the skewness errors that the aforementioned columns provided. Because of this, Ratings and Votes have become redundant columns that disobey Normal Form 3 and have to be removed.

Budget

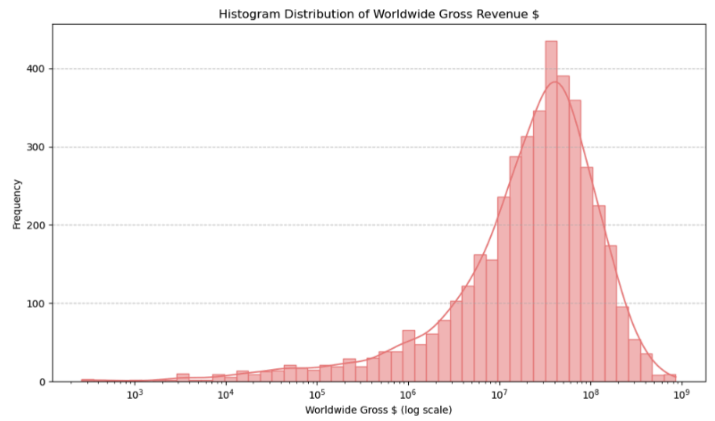


Budget is an important feature that may have high predicting value for our target variable. The distribution is **skewed to the right** with **many, yet important, outliers** that still abide by the relationship that the higher the budget, the higher the gross revenue. Hence, scaling would be more appropriate than removing these rows.



Domestic Gross $

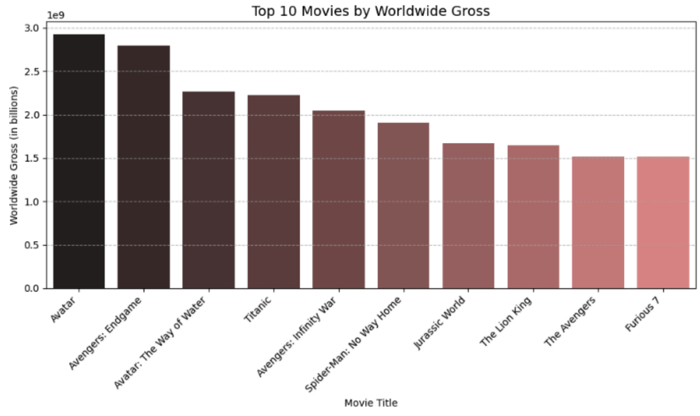
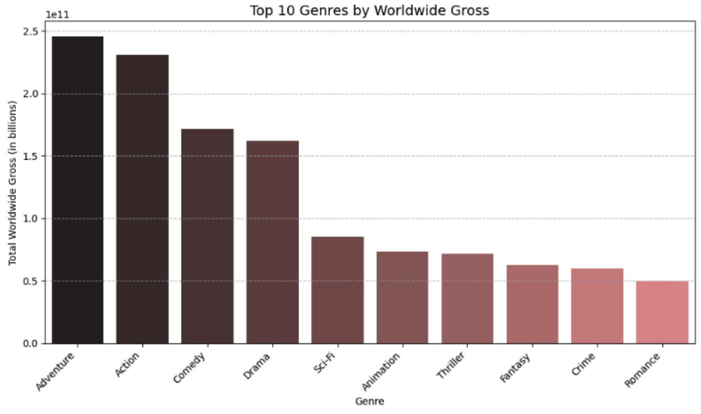
Takes on a normal distribution that is slightly skewed to the right.

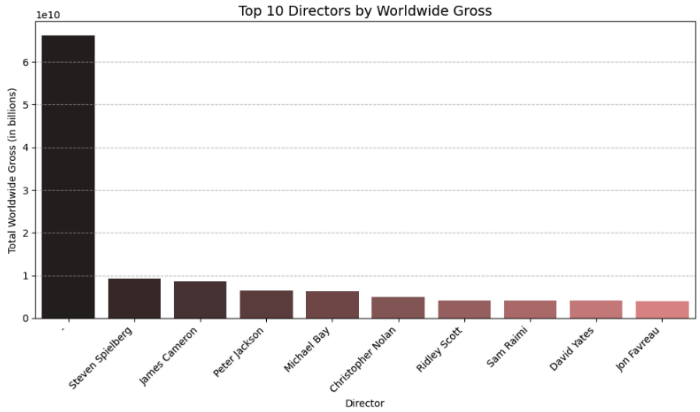
	**Worldwide Gross $** Takes on a normal distribution that is slightly skewed to the right.

2. **Review of Frequency Outliers**

   Due to the complexity of the dataset, there are a lot of outliers present since blockbuster movies often achieve success due to factors beyond predictable trends. However, important correlations still exist—such as the positive relationship between budget and gross revenue—and these outliers generally follow this pattern. Hence, it is important to include them. To account for these outliers, I will be using tree-based algorithms, which are more robust to such anomalies, and will apply scaling and binning techniques where appropriate rather than omitting the data.

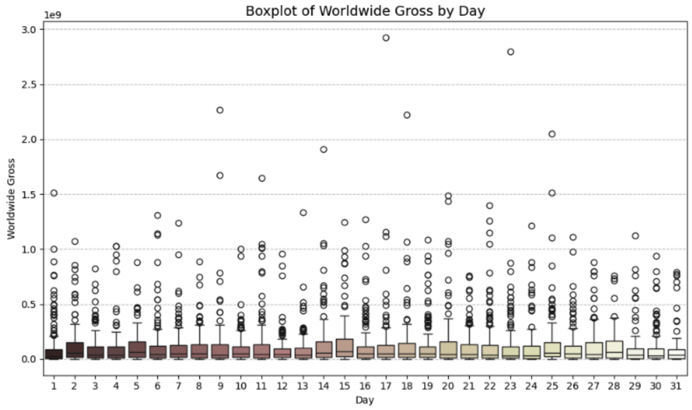
3. **Comparison between Feature and Target Variable, Worldwide Gross $**

	**Top 10 Movies** Movie titles are generally irrelevant to how well a movie will perform.
	**Top 10 Genres** Genres such as Adventure and Action may give way to more successful movies due to their popularity.



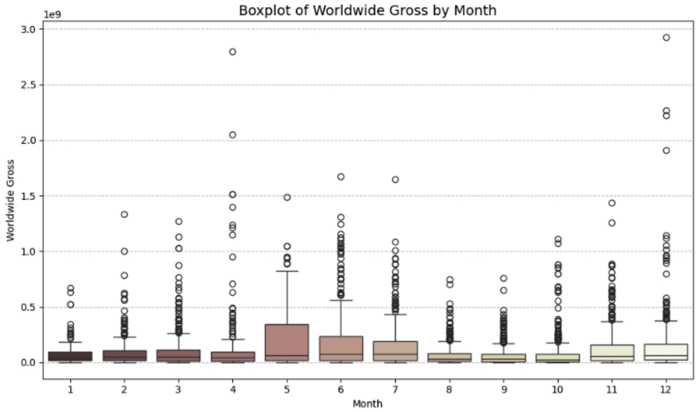
Top 10 Directors

Like our observation above, there are too many movies that are not associated with a director hence, this incomplete feature should not be included in the model-building process.



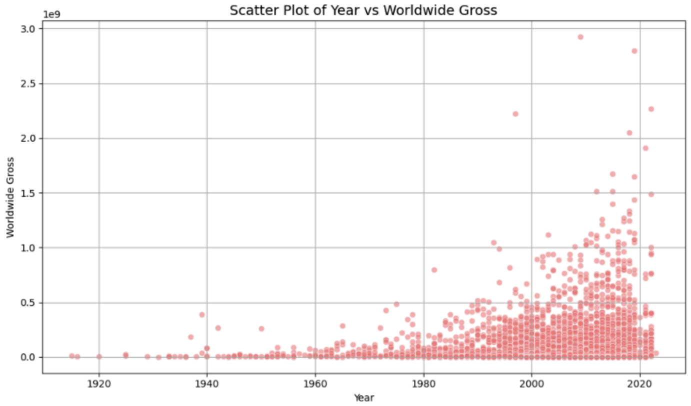
Worldwide Gross $ by Day

From the boxplot diagram, the day the movie was produced does not really impact how well it will perform.



Worldwide Gross $ by Month

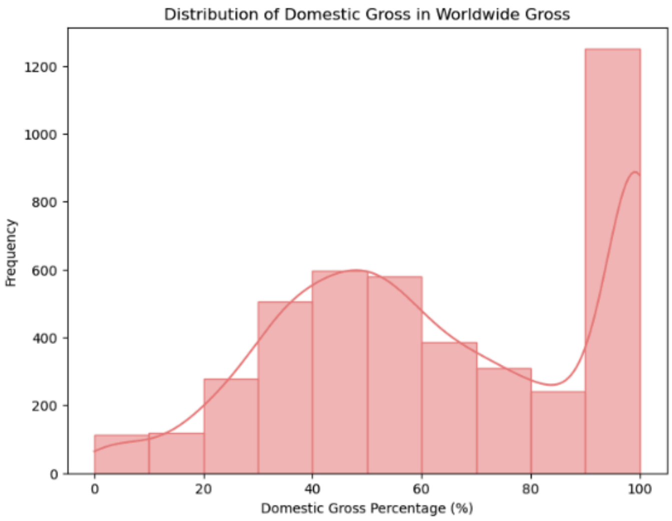
Like the Day, based on the box plot diagram, the month the movie was produced does not really impact how well it will perform. There are several outliers, and their distribution is unique but may produce noise in our model-building process instead of aiding.



Worldwide Gross $ by Year

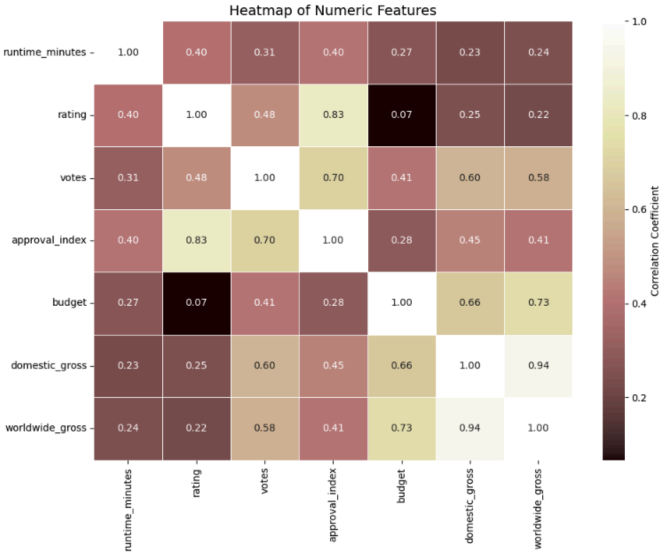
From the scatter plot, there is a relationship between the year and worldwide gross, affirming my aforementioned predicted relationship.

4. Relationships



Distribution of Domestic Gross $ in Worldwide Gross $

From this histogram, a number of movies have 90-100% of their worldwide gross $ consisting of the domestic gross $. This poses as a **data leakage threat**, especially when domestic gross = worldwide gross, having a **100% correlation**. As such, this feature needs to be removed.

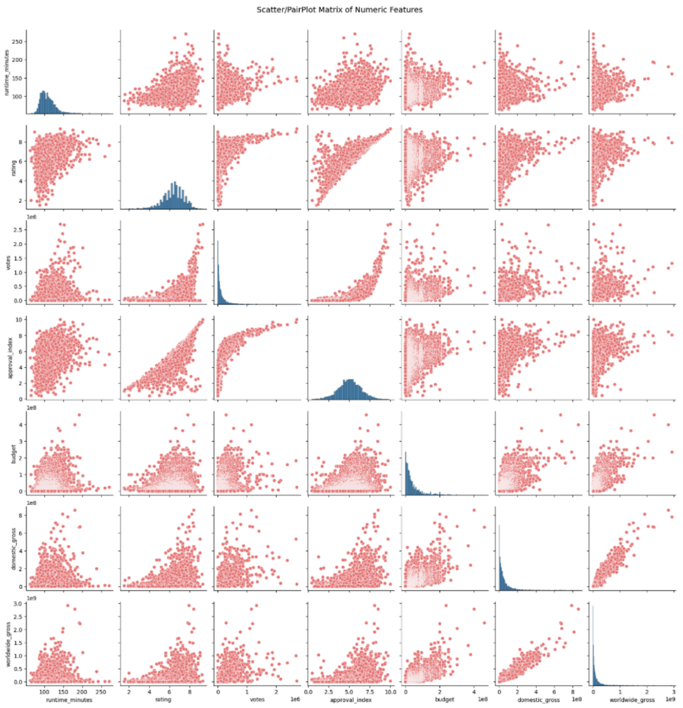


	runtime_minutes	rating	votes	approval_index	budget	domestic_gross	worldwide_gross
runtime_minutes	1.00	0.40	0.31	0.40	0.27	0.23	0.24
rating	0.40	1.00	0.48	0.83	0.07	0.25	0.22
votes	0.31	0.48	1.00	0.70	0.41	0.60	0.58
approval_index	0.40	0.83	0.70	1.00	0.28	0.45	0.41
budget	0.27	0.07	0.41	0.28	1.00	0.66	0.73
domestic_gross	0.23	0.25	0.60	0.45	0.66	1.00	0.94
worldwide_gross	0.24	0.22	0.58	0.41	0.73	0.94	1.00

Heatmap Correlation of Numeric Variables

Notable Relationships:

- Worldwide Gross $ and Budget
- Rating, Votes, and Approval Index
- Domestic Gross $ and Runtime Gross $



Pairplot Matrix of Numeric Variables

From the plots, runtime minutes, approval index, and budget are features that would be helpful in the model building process due to having more noticeable relationships.

2.3 Data Cleaning

Handling missing values, correcting errors, dealing with duplicates, and removing outliers to ensure high-quality data.

1. **Removing Genre Outlier**
   As mentioned in the beginning, it is important to remove this particular outlier due to it's insignificance in the result as well as its potential to negatively impact our model. After any manipulation in the dataframe, a reordering of the index is necessary to avoid complications in the future.

2. **Removing Unnecessary Columns**
   The movie_title column is removed as it does not contribute to predicting worldwide gross revenue and lacks predictive value.

3. **Removing Columns with Unclean Values**
   Columns containing irrelevant or incomplete data, such as the director's personal information, are removed. Many of these columns are missing data or contain fillers such as '-' or '\N', making them imbalanced and unsuitable for prediction.

4. **Feature Splitting - Splitting production_date into 3 Separate Columns**
   The year-month-day format is not ideal according to data field normalisation principles, First Normal Form. The year column is separated; it is a key factor influencing worldwide gross prediction, while the day and month are not as impactful, as seen in the above visuals, and are therefore excluded.

5. **Removing Redundant Columns**
   Since approval_index is a result of the combination of rating and votes, these two columns are removed to maintain data integrity and adhere to Third Normal Form principles.

6. **Target Leakage Variables**
   Due to many rows in domestic gross $ having a 100% correlation with worldwide gross $, it is removed as it could unfairly influence the model's predictions.

2.4 Data Pre-Processing

The preparation of raw data for analysis or modelling by **transforming** it into a format that machine learning algorithms can work with. This includes normalising or scaling the data to ensure consistency across features, encoding categorical variables into numerical forms since machine learning models require numeric input, and incorporating methods to account for outliers, such as discretisation. However, the methods chosen must be carefully considered as many normalisation methods use the entire dataset so these can only be conducted after the dataframe has been split into test and train. Otherwise, these models would use unseen data thus resulting in **data leakage**.

1. **Discretisation - Binning Runtime Minutes**
   The process of converting continuous data into discrete categories or bins. Due to the large range of movie runtime minutes, it is useful to group these values into specific ranges to reduce complexity, improve performance, and make the data more interpretable. Binning is a common method for this as it helps simplify the data by reducing the number of unique values, which can be particularly useful for features that have a large range or skewed distributions, such as movie runtime minutes. I implemented a custom bin range based on the quartiles of the feature that we have identified during profiling. Since the majority of movies fall under the 90-120 window, this will be Medium with the bins following accordingly:

Short	< 90 Minutes
Medium	90 - 150 Minutes
Long	150 - 210 Minutes
Very Long	> 210 Minutes

   Benefits:
   a. Simplifies the analysis - Makes relationships in the data more apparent.
   b. Handles outliers - Extreme values (Abnormally short or long movies) may not need to be treated as outliers as binning helps group them into a meaningful category without distorting the model.
   c. Improve model performance - This can lead to a more robust model by reducing the noise and making the data easier to work with.

2.5 Feature Engineering

Creating new features, transforming existing ones, and selecting the most relevant features to improve model performance.

2.5.1 Before Splitting and Shuffling

1. **One-Hot Encoding Runtime Minutes**
   A method used to convert categorical data into a numerical format, which is a necessary step because machine learning algorithms typically require numeric input. It creates binary columns for each bin that we have made above, containing a value of 1 if the instance belongs to that bin, or 0 otherwise.

   Benefits:
   a. Categorical Representation - Easier to represent runtime data and makes it useful because decision trees require the numeric form
   b. Avoiding Ordinal Assumptions - This process treats these bins as independent categories, which is useful because there is no ranking between them and prevents multicollinearity

2.5.2 After Splitting and Shuffling

1. **Normalisation - Robust Scaling Budget**
   A technique used to scale features in a dataset, especially when outliers are present. This method uses the median and interquartile range to handle the influences of extreme values more effectively. By using this method, it ensures that all features are on the same scale and that outliers do not have a disproportionate influence. Since the budget data is expected to contain high-budget blockbuster movies that can significantly skew the data, robust scaling ensures that these extreme values don't distort the scaling of the other values in the dataset.

$$x_{scaled} = \frac{x - median(x)}{IQR(x)}$$

   Benefits:
   a. Immune to outliers
   b. Can improve machine learning algorithm performance
   c. More stable, hence less likely to be affected by changes in the data
   d. More robust, hence less likely to overfit the training data

   Comparison to Log Transformation:
   On the other hand, this method, while commonly used, can still be affected by outliers as in most cases, it does not help make data more normal, and in some circumstances, made data more skewed.

2. **Normalisation - Standard Scaling Year**

Standard Scaling, also known as Z-Score normalisation, is a method used to standardise the features of a dataset so that they have a mean of 0 and a standard deviation of 1. This is particularly useful as it is less sensitive to outliers, unlike min-max scaling, which is what appears to be the case for Year, which tend to fall within a relatively limited range.

$$z = \frac{x - \mu}{\sigma}$$

$x - original\ value$
$\mu - mean$
$\sigma - standard\ deviation$

Benefits:

a. Preserves distribution shape
b. Enables faster convergence - Models train faster due to the uniform input ranges

3. **Dimensionality Reduction - Principle Component Analysis**

A technique used to reduce features in a dataset while retaining the essential information. It helps simplify the dataset, making it more manageable for modeling, and can improve the performance of machine learning algorithms by removing noise and redundancy.

PCA works by identifying the directions (principal components) in which the data varies the most. These principal components are new, uncorrelated features that are linear combinations of the original features. The first principal component captures the most variance, the second captures the second most, and so on. By selecting the top principal components, PCA reduces the number of features while preserving as much information as possible.

Necessity:

1. After conducting one-hot encoding on the genres, I was left with 74 additional features, one column for one genre. Due to the curse of dimensionality, this high-dimensional data can lead to issues like overfitting, where the model becomes too complex and does not generalise well.
2. Many of these genres are likely to be correlated with each other, leading to multicollinearity, which can negatively impact the model’s performance and interpretation.
3. By reducing dimensionality, PCA helps in filtering out noise from less important genres, ensuring that the model is focused on more relevant patterns.

PCA must be conducted **after** the train-test split as this process takes the entire dataset into consideration hence may result in data leakage.

3.0 Methods and Improvements

3.1 Parameter Tuning

Iterations **1000 to 1500** 1000	Specifies the number of boosting iterations (or trees) that will be used during the training of the model. More iterations generally lead to better fitting, but can also cause overfitting if set too high.
learning_rate **0.01 to 0.1** 0.03	Controls how quickly the model adapts to the problem. A lower learning rate means the model will learn slowly but might be more accurate with more iterations.
depth **6 to 10** 6	Controls the maximum depth of the trees. Deeper trees can model more complex relationships but are more prone to overfitting. Shallow trees are faster to train but may underfit.
l2_leaf_reg **1 to 10** 3	A regularisation term that helps to reduce overfitting. It penalises large leaf values and encourages simpler models.
bagging_temperature **0 to 3** 1	Controls the degree of randomness in the selection of training data during each iteration. Higher values make the model more robust to noise by reducing the chance of overfitting, but it can slow down learning.
border_count **200 to 255** 254	Defines how the model will divide continuous features into discrete bins.
subsample **0.7 to 0.9**	Controls the fraction of the dataset that will be used for each tree. This can help with preventing overfitting by introducing randomness during training.
random_strength **1 to 3**	Controls the randomness introduced when splitting trees. It can be used to regularise the model and prevent overfitting.

3.2 Checks and Considerations

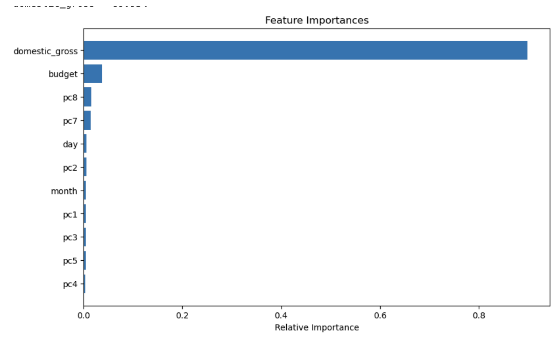
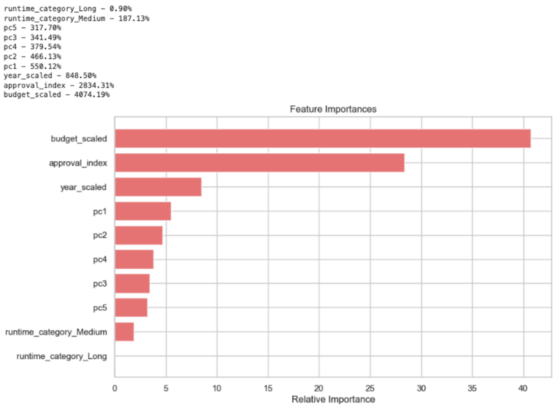
1. **Feature Multicollinearity - Variance Inflation Factor (VIF) Function**
   Feature multicollinearity refers to a situation where two or more features are highly correlated with each other. It then becomes difficult for the model to determine the individual effect of each correlated feature. The Variance Inflation Factor (VIF) is a statistical measure that quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other features. A high VIF indicates that a feature is highly correlated with other features, leading to multicollinearity. Typically, a VIF above 10 suggests problematic multicollinearity.

2. **Data Leakage Identification Function**
   Occurs when information from outside the training dataset is used to create the model, for example, information that isn’t available during the time of prediction. This can lead to over-optimistic results during model training because the model has access to information it wouldn't have in a real-world prediction scenario. It typically occurs when the target variable leaks into the features used to predict the target.

3. **Feature Importances Function**
   This refers to the process of identifying which features have the most significant impact on the prediction of the target variable. It helps identify which features should be kept in the model and which might be irrelevant or redundant.

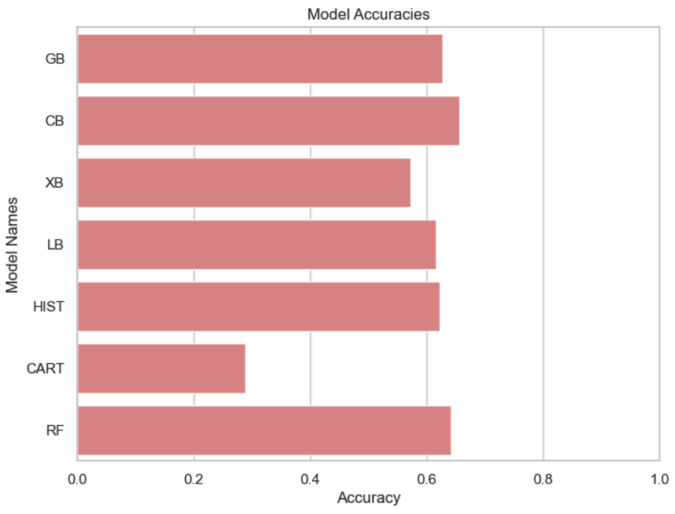
3.3 Mistakes and Reflection

Conducting Over-Pre-Processing
Originally, I applied almost all pre-processing techniques, thinking that it would enhance the dataset to the fullest extent. However, this approach led to unintended data leakage, which compromised the integrity of the model. I realised that quality > quantity to ensure that my model was truly the best I could produce.

Before	After
	runtime_category_Long - 0.90% runtime_category_Medium - 187.13% pc5 - 317.70% pc3 - 341.49% pc4 - 379.54% pc2 - 466.13% pc1 - 550.12% year_scaled - 848.50% approval_index - 2834.31% budget_scaled - 4074.19%

4.0 Results and Analysis

Model Accuracies Bar Chart with Default Parameters



GradientBoost: **0.62**7660494032448

CatBoost: **0.65**60338115390995

XGBoost: **0.57**26083517074585

LightBoost: **0.61**63782329697894

HistogramBoost: **0.62**23349469186334

CatAndRegTree: **0.28**81859537420336

RandomForest: **0.64**24783701484875

Model of Choice: Category Boosting

A powerful gradient boosting algorithm specifically designed to handle categorical features efficiently. It is good for machine learning tasks involving structured data by offering robust handling of categorical variables, reducing overfitting, and improving training speed. Unlike traditional boosting algorithms that require extensive preprocessing to convert categorical data into numerical form (e.g., one-hot encoding), CatBoost natively supports categorical features, automatically processing them using advanced encoding techniques such as ordered boosting and target-based statistics, which prevent data leakage and overfitting.

It's appropriate due to its robustness in handling categorical features like movie genres and production years without requiring excessive preprocessing steps, thereby improving both efficiency and model accuracy.

Statistical Analysis

Best Parameters	Iterations: 1840 Learning_rate: 0.02265134588264378 Depth: 6 L2_leaf_reg: 9.630098170118737 Bagging_temperature: 0.4420592524429963 Border_count: 238 Subsample: 0.7202445085190277 Random_strength: 2
Run Time	1191.713s
Mean Absolute Error on **TRAIN**	41232740.02945805
Mean Absolute Error on TEST	56412708.12521616
Mean Squared Error on TEST	1.0647195803197016e+16

Root Mean Squared Error on TEST	103185249.93038984
R-Squared on TEST	0.6691263751606404
Adjusted R-Squared on TEST	0.6670980311003225
Adjusted R-Squared on **TRAIN**	0.668259930316544

Since the worldwide gross revenue values are in the billions, the Mean Absolute Error (MAE) naturally appears high. However, what truly matters is the relative difference between the training and testing MAE values. In this case, the difference between the two is not substantial, indicating that the model is not overfitting and is generalising well to unseen data.
Similarly, the R-squared (R^2) and adjusted R-squared values, while not perfect, suggest a reasonable fit. The close values between the training and test sets further confirm that the model captures a significant portion of the variance in revenue without overfitting. This balance demonstrates that the model's performance is consistent and reliable, despite the high absolute error values.

5.0 Conclusion

Final Model: Category Boost

Most Influential Features:

Budget	A higher budget enables superior production quality, extensive marketing, and global distribution, directly contributing to higher worldwide gross revenue. It allows for star-studded casts, advanced visual effects, and broader audience appeal, leading to increased ticket sales.
Approval Index	Positive audience and critic reviews boost a movie's reputation, driving higher attendance through word-of-mouth and prolonged theatre runs. A strong approval index often correlates with greater revenue due to increased viewer trust and repeat viewings.
Production Year	The year of production influences a film's revenue due to evolving audience preferences, economic conditions, and technological advancements. Market trends and external factors like the COVID-19 pandemic also impact box office performance.
Genre of Movie	Certain genres, such as action and adventure, have widespread global appeal, attracting larger audiences and generating higher revenue. In contrast, niche genres may have limited market reach, affecting their earning potential.
Movie Duration (Runtime Minutes)	The length of a film affects both audience engagement and theatre scheduling. Moderately timed films tend to perform better, balancing storytelling with viewer attention spans, while excessively long runtimes may reduce screening opportunities.

6.0 References

- *What is different between EDA, Data Preprocessing, Feature Engineering?* | Kaggle. (2025). Kaggle.com. https://www.kaggle.com/discussions/getting-started/422362
- Hao, B. (2023). The Analysis of the Factors that Influence the Film Revenue. *Highlights in Science Engineering and Technology, 47,* 154–159. https://doi.org/10.54097/hset.v47i.8184
- Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *PubMed*. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009
- GeeksforGeeks. (2024, February 28). *Tree Based Machine Learning Algorithms.* GeeksforGeeks. https://www.geeksforgeeks.org/tree-based-machine-learning-algorithms/
- Raghav Vashisht. (2021, January 6). *Machine Learning: When to perform a Feature Scaling? - Atoti Community. Atoti Community.* https://www.atoti.io/articles/when-to-perform-a-feature-scaling/
- Islam, N. (2023, November 28). *Mastering Exploratory Data Analysis (EDA): A Comprehensive Python (Pandas) Guide for Data Insights and Storytelling*. Medium. https://medium.com/@nomannayeem/mastering-exploratory-data-analysis-eda-a-comprehensive-python-pandas-guide-for-data-insights-c0be7c5b8889
- *Robust Scaling: Why and How to Use It to Handle Outliers | Proclus Academy.* (2022, March 22). @_yashmeet. https://proclusacademy.com/blog/robust-scaler-outliers/
- *Movie Revenue Prediction Using Machine Learning Models.* (2020). Arxiv.org. https://arxiv.org/html/2405.11651v1
- *How To Find Outliers Using Python [Step-by-Step Guide].* (2022, February 21). CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/
- *Minitab Blog Editor. (2025). Enough Is Enough! Handling Multicollinearity in Regression Analysis.* Minitab.com. https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis
- *How to choose the right data visualization chart type - 2025 Guide | Zoho Analytics.* (2024, November 6). Zoho Analytics. https://www.zoho.com/analytics/insightshq/choosing-the-right-data-visualization-type.html
- Team, E. (2023, January 2). *What is Data Analytics? Definition, Types, case study, and more.* Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/what-is-data-analytics/