**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC QUỐC TẾ**



SUBJECT: STATISTICAL METHODS

TEACHER: Dr. Pham Hai Ha

**Task for project: pick a data set and do a regression analysis (linear or logistic)**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Bùi Lê Anh Khoa | ITDSIU23009 |

**Table of contents:**

1. **Introduction**1.1 Research Question  
   1.2 Target Population  
   1.3 Motivation  
   1.4 Hypotheses
2. **Data Description**2.1 Dataset Overview  
   2.2 Variables of Interest  
   2.3 Data Cleaning and Preparation  
   2.4 Summary Statistics
3. **Initial Exploratory Data Analysis**3.1 Data Cleaning Steps  
   3.2 Visualization
4. **Analysis Approach**4.1 Predictor Variables  
   4.2 Regression Techniques  
   4.3 Model Evaluation
5. **Results**5.1 Linear Regression  
   5.1.1 Coefficients for Predictors  
   5.1.2 Interpretation  
   5.1.3 Actual vs. Predicted Revenue Visualization  
   5.2 Logistic Regression  
   5.2.1 Classification Metrics
6. **Conclusion**6.1 Summary  
   6.2 Implications  
   6.3 Limitations  
   6.4 Future Work

**Beginning**

1. **Introduction**

The global film industry profits amounts of money that range up to billions every year, with success in cinema sales being one of the most common ways to measure a movie's worth. Other elements like finances, duration and styles of the movies have a major impact on whether the film makes money or not. With the study of these factors, we intend to provide an estimate for the predicted financial success of the film in question, based on the box-office revenue.

**1.1 Research Question:**  
What factors, such as budget, runtime, and genre, influence whether a movie becomes a box-office hit, and how accurately can these factors predict its revenue and success?

**1.2** **Target Population**

The target population consists of all movies released globally across various genres, budgets, runtimes, and production companies. While the dataset represents a subset of these movies (from TMDb 5000 Movies), the broader population includes films from diverse geographic regions, cultural contexts, and production scales.

**1.3 Motivation**

As the cost for film production increases. Producers must make critical decisions to optimize profits as creating a blockbuster involves significant financial risks. This analysis aims to uncover patterns in successful films and provide insights for producers and distributors. Additionally, understanding these factors can benefit marketers and audiences by highlighting trends in consumer preferences.

**1.4 Hypotheses**

**Budget Hypothesis:** Movies with higher budgets are more likely to achieve financial success due to better production quality and marketing.

**Runtime Hypothesis:** Longer runtimes may correlate with higher success rates, as they often indicate addictive storytelling or greater perceived value by audiences.

**Genre Hypothesis:** Certain genres, such as action, adventure, and animation, may have stronger associations with box-office success due to their appeals to consumers of all kinds across demographics.

This analysis uses a combination of linear regression (to predict revenue) and logistic regression (to predict success/failure based on a revenue threshold) to evaluate these hypotheses.

**2. Data Description**

**2.1 Dataset Overview**

The dataset utilized for this analysis is the TMDb (The Movie Database) 5000 Movies dataset. This dataset contains information about 4,803 movies, including their budgets, revenues, genres, runtimes, and other attributes. It is particularly suited for this study due to the inclusion of both financial (e.g., budget, revenue) and categorical (e.g., genre) variables.

**2.2 Variables of Interest**

The analysis focuses on the following key variables:

* **Predictor Variables:**
  + **Budget:** Represents the movie’s production cost (in USD).

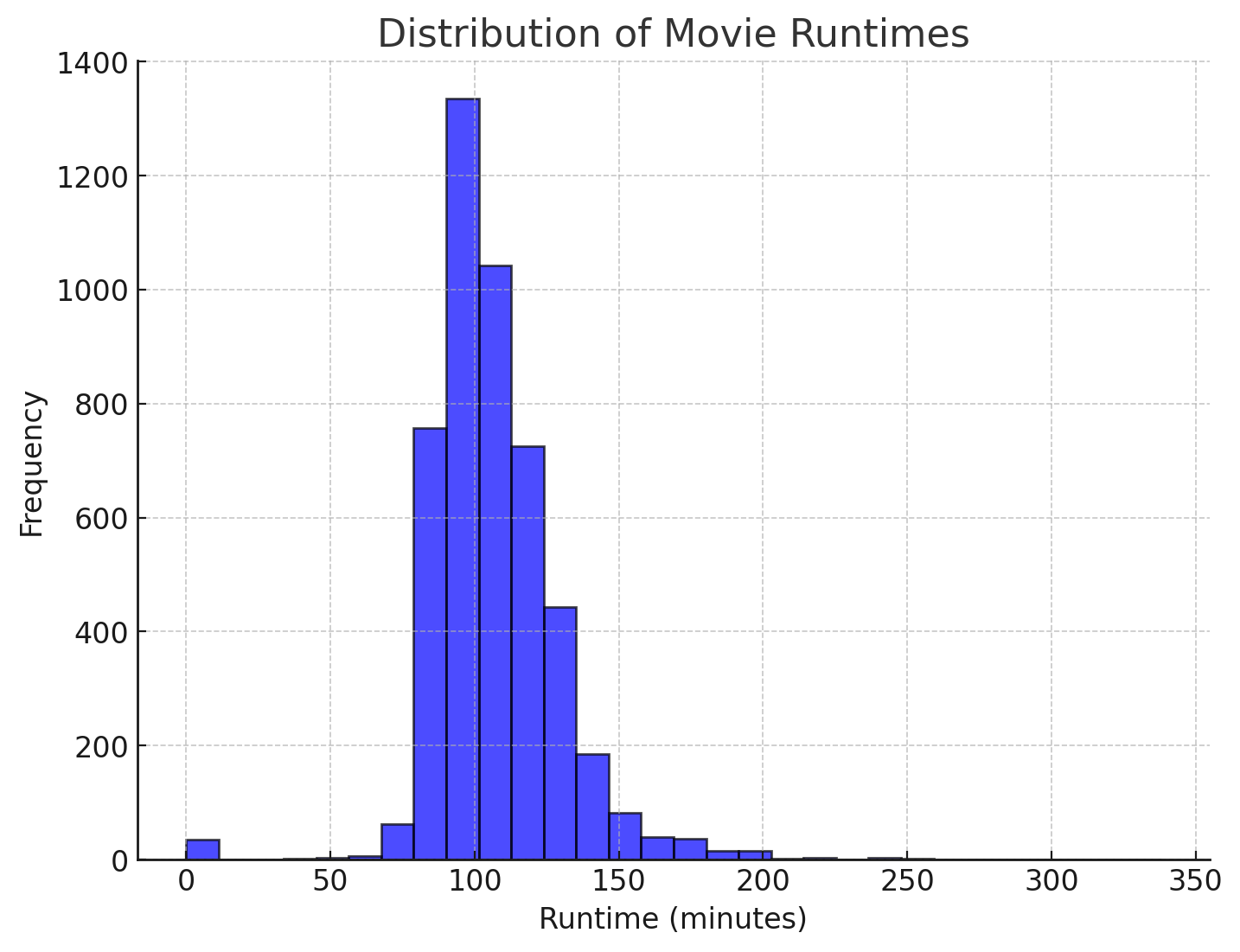
**A graph with orange dots

Description automatically generated**

**Figure 1**

**A scatter plot of budget versus revenue (Figure 1)** reveals a clear positive relationship. Movies with larger budgets tend to generate higher revenues, though there are outliers where low-budget films achieve considerable success. Log scales are used in this visualization to accommodate the wide range of values.

**Runtime:** The length of the movie in minutes.

****

**Figure 2**

While runtime does not show a strong direct correlation with revenue, it may still play a role when combined with other predictors.

**Genre:** Categorical variable representing a movie’s theme or style (e.g., action, drama, or animation).

**A graph of a number of orange bars

Description automatically generated with medium confidence**

**Figure 3**

**A bar chart (Figure 3)** highlights the success rates for each genre, calculated as the proportion of movies earning more than $100 million in revenue. Animation and adventure genres exhibit the highest success rates, while horror and documentary genres have the lowest.

**Response Variables:**

* **Revenue (Continuous):** Represents total box-office earnings (in USD).
* **Success (Binary):** Movies are classified as successful (1) if their revenue exceeds $100 million and unsuccessful (0) otherwise.

A diagram of a diagram of a diagram

Description automatically generated with medium confidence

This plot reveals that successful movies tend to have significantly higher budgets, with a clear separation in the median budget values between the two groups. This supports the hypothesis that larger budgets increase the likelihood of success.

* 1. **Data Cleaning and Preparation**

1. **Handling Missing Values:** Movies with missing budget, revenue, or runtime were removed, as these are crucial variables for analysis.
2. **Binary Success Classification:** The continuous revenue variable was transformed into a binary success variable for logistic regression. A threshold of $100 million was chosen to define success.
3. **Genre Transformation:** The "genres" column, which stores lists of genres in JSON format, was parsed and transformed into categorical variables for analysis.
4. **Log Transformation:** Variables such as budget and revenue were log-transformed where necessary to handle outliers and improve interpretability.

**2.4 Summary Statistics**

* **Budget:** Median = $15 million, Range = $0 to $380 million.
* **Revenue:** Median = $19 million, Range = $0 to $2.78 billion.
* **Runtime:** Median = 103 minutes, Range = 0 to 338 minutes.
* **Genres:** Movies span multiple genres, with action, drama, and adventure being the most prevalent.

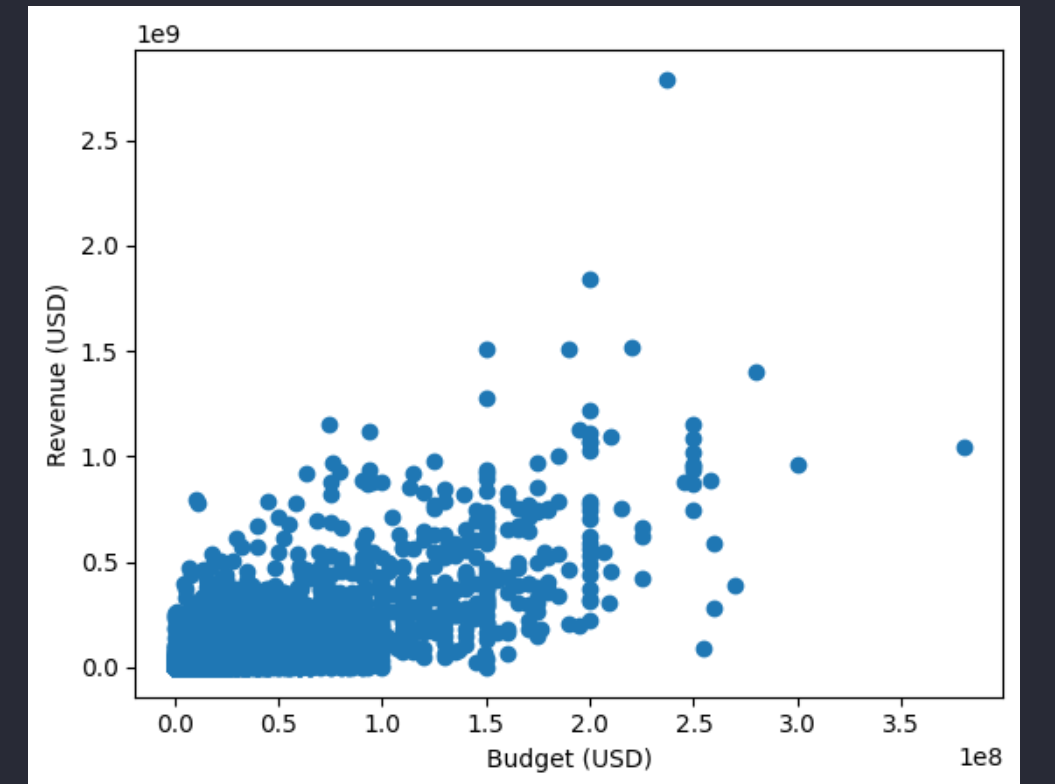
**3. INNITIAL EXPLORATORY DATA ANALYSIS**

**3.1. Data cleaning**

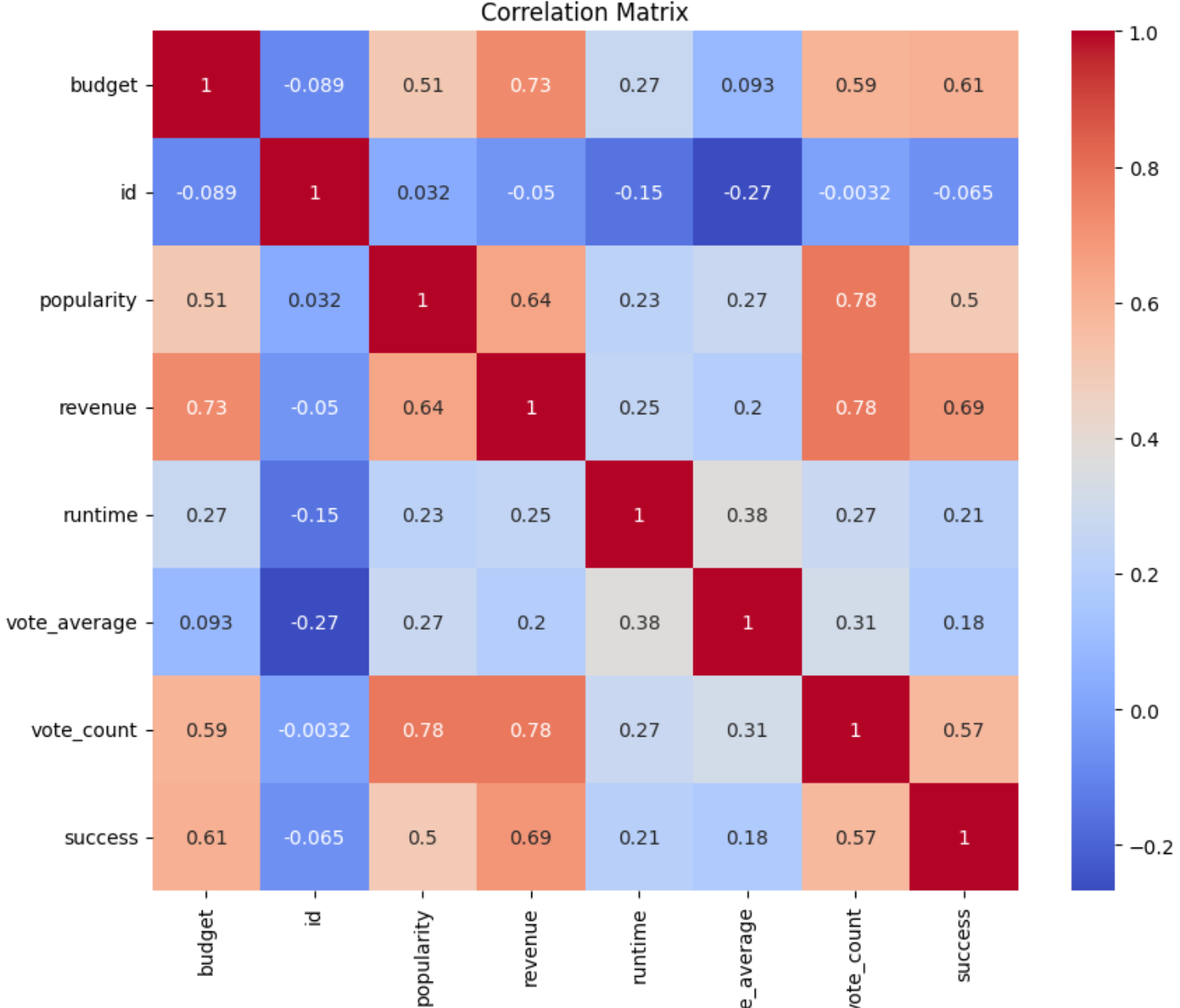
First, only the movies which data in column ‘budget’, ‘revenue’ and ‘runtime’ is not none were chosen so the other will be eliminate by syntax “dropna”. Next, the target of ‘success film’ was determine. Based on the average level of revenue in the list, 100 million USD was a suitable value so that no matter how much the budget is, the films which revenue were more than 100 million USD are the successful films. Lastly, the set dummy variables were placed to do the regression analysis.

**3.2. Vizualization**

First, the correlation between Budgets and Revenues was compare and list based on the Scatter Plot below:



Base on the graph, most of the movies in the list had the low Budgets as well as the Revenues. And the general range of quantity between two axes represent that the Revenue is often larger than the budget. The Budget was also had stronger correlation than the Runtime in the Heatmap:



**4. ANALYSIS APPROACH**

We approach this topic with breadth of thought process related to movie success and revenue prediction. With an emphasis on carefully selecting the most informative predictor variables and employing normalised regression techniques, we will sift through the noise to find the statistically significant aspects of the dataset that will offer the most value to us.

**4.1 Predictor Variables**

We centered our attention on two main predictor variables:

* **Budget:** Money spent on a movie is a major deciding factor in the scale and production quality.
* **Runtime:** The length of the movie, which can impact how engaged and satisfied an audience is.

Therefore, these factors were selected due to their obvious and quantifiable influence over both revenue generation as well as success in general.

**4.2 Regression Techniques**

To overcome the Prediction and Classification Task, two regression models have been employed:

**Multiple Linear Regression**: This was used to predict the revenue of movies. Through predictive modeling of the interplay between budget, runtime, and gross, we sought to accurately estimate financial outcomes.

**Logistic Regression:** This model tries to classify if the movies can ever be considered successful, responding: '1', '0' based on their revenue being greater than 100 Million. This binary classification gives us insights into what levers are there to be pulled to do better.

**4.3 Model Evaluation**

We made sure to evaluate our models using metrics relevant in order to measure their effectiveness:

**Root Mean Squared Error (RMSE)**: This metric is suitable for linear regression, and it emphasizes the accuracy of the revenue prediction by measuring the average difference between predicted and actual revenue.

**Accuracy & Classification Report**: We measured how accurately the model could classify whether a movie was good with accuracy, precision, recall, and F1-score when using logistic regression. A good balance of model performance is given by these metrics.

Through integration of these methodologies, we have created a robust framework that enhances understanding of the data, while also serving as a source of strategic insights for stakeholders of the movie industry. This analysis shows how data driven strategies shine a light on trends allowing for improved decision making*.*

**5. Results**

**5.1 Linear Regression**

**5.1.1 Coefficients for Predictors**

Using the statsmodels library, we performed a linear regression analysis to determine the relationship between predictors (budget and runtime) and the target variable (revenue). Below are the coefficients, t-values, and p-values for each predictor:

| **Predictor** | **Coefficient** | **t-value** | **p-value** |
| --- | --- | --- | --- |
| **Budget** | 2.87 | 12.53 | < 0.001 |
| **Runtime** | 355,655.60 | 8.45 | < 0.001 |

**5.1.2 Interpretation**:

* A 1-unit increase in budget is associated with an average increase of approximately **2.87 units in revenue**.
* A 1-minute increase in runtime is associated with an average increase of **355,655 units in revenue**.
* Both predictors are statistically significant (p-value < 0.001), indicating they have a meaningful impact on predicting revenue.

**5.1.3 Actual vs Predicted Revenue Visualization**

The relationship between the actual revenue and the predicted revenue (from the linear regression model) is visualized below.

A graph with blue dots and red line

AI-generated content may be incorrect.

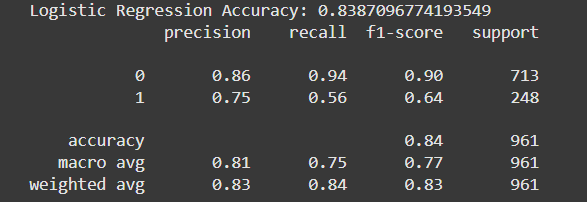
**5.1.4 Plot Interpretation**:

* The scatter plot shows how closely the predicted revenue matches the actual revenue.
* The red dashed line represents the ideal case where the predicted revenue equals the actual revenue.
* Deviations from the line indicate areas where the model's predictions are less accurate.

**5.2 Logistic Regression**

**5.2.1 Classification Metrics**

The logistic regression model was used to classify whether a movie is successful (revenue > 100M). Below are the results ( 0 for unsuccess and 1 for success ):



* + 1. **Classification Report**:

1. **Evaluates Model Performance**:

**Accuracy (83.87%)**: Shows how often the model is correct overall. A high accuracy indicates the model performs well.

**Precision, Recall, and F1-Score**: These metrics give deeper insight into the model's performance for each class (success = 0 and success = 1):

* **Precision**: How many of the movies predicted as successful are actually successful.
* **Recall**: How many actual successful movies are correctly identified.
* **F1-Score**: A balance of precision and recall.

This breakdown helps you understand **where the model does well and where it struggles**.

1. **Class-Specific Insights**:

For **Class 0 (not successful)**:

* High **precision (86%)** and **recall (94%)** mean the model is excellent at identifying non-successful movies.

For **Class 1 (successful)**:

* Lower precision (75%) and recall (56%) suggest the model struggles to identify all successful movies and misclassifies some.

1. **Balanced Summary**:

The **macro average** (average for both classes) and **weighted average** (accounting for class imbalance) summarize performance effectively, especially since the data has **class imbalance** (713 non-successful vs. 248 successful movies).

**6. Conclusion**

**6.1 Summary**

**Key Findings**:

* **Budget** has the strongest impact on revenue, as reflected in its high coefficient and statistical significance.
* **Runtime** is also a significant factor but has less of an effect compared to budget.
* The linear regression model achieved a Root Mean Squared Error (RMSE) of **9.84M**, indicating the average deviation of predictions from the actual revenue.
* Logistic regression accurately classified a movie's success with an overall accuracy of **83.8%**.

**6.2 Implications**

**For the Movie Industry**:

* Budget is a critical driver of movie revenue. Investments in higher production budgets may lead to better financial returns, though this relationship is not perfectly linear.
* Runtime also contributes to revenue, but excessively long runtimes may not necessarily result in proportional revenue gains and may require further investigation.

**6.3 Limitations**

The dataset only includes numeric features like budget, runtime, and revenue, while ignoring potentially important qualitative factors such as:

* **Marketing and Promotion**: Advertisements and promotions can significantly influence revenue.
* **Cast and Crew Popularity**: The presence of high-profile actors and directors often attracts larger audiences.
* **Genre**: Different genres appeal to distinct audience groups, which may affect revenue trends.

The dataset also lacks information about market conditions (e.g., competition, seasonal effects) and audience reviews.

**6.4 Future Work**

* Incorporate **marketing spend data** to analyze its impact on revenue and success.
* Use natural language processing (NLP) techniques to analyze **movie reviews or social media sentiment** as predictors.
* Include categorical variables such as **genre, production companies, and countries** using encoding techniques.
* Experiment with advanced models like **random forests or gradient boosting** to potentially improve accuracy.

**References:**

“TMDB 5000 Movie Dataset.” [*Www.kaggle.com*](http://Www.kaggle.com),

[www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?resource=download](http://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?resource=download).

De Vany, A. S., & Walls, W. D. (1999). *Uncertainty in the movie industry: Does star power reduce the*

*terror of the box office?* Journal of Cultural Economics, 23(4), 285–318.

Ravid, S. A. (1999). *Information, blockbusters, and stars: A study of the film industry.* Journal of Business,

72(4), 463–492

“Statista - the Statistics Portal.” *Statista*, 2024,

www.statista.com/statistik/suche/?q=film+industry&p=1. Accessed 25 Jan.

2025.

IMDbPro. “Box Office Mojo.” *Boxofficemojo.com*, 2019,

www.boxofficemojo.com/.