Movie Project Outline

## Project Introduction

Investigate how different movie attributes (genre, plot, acting, pacing, cinematography, soundtrack) influence review scores and box office performance.

**Significance:** Helps studios and marketers predict success and optimize movie production.

**Research Questions:**

1. How do specific movie attributes (genre, acting, cinematography, soundtrack, runtime, director, writer, star power) correlate with review scores and box office revenue? Which attributes are most important for each types of movie genre?
2. How do certain actors, writers, directors, run times, and genres impact review score and box office performance?
3. Do higher review scores lead to higher box office revenue?
4. Predicting Box Office Revenue Using Review Sentiment and Movie Attributes

## Data sets Used

1. imdb data set: 52452 variables. Noticably, movie id, movie name, genre, runtime, rating, director, director id, star, star id, box office
2. Rotten Tomatoes Movie Reviews: 1444963 variables. movieid, review id, critic, is topcritic, originalscore, review text
3. Rotten Tomatoes Movie: 143258 variables. Movieid, audience score, tomato meter, runtime, genre, director, writer, box office

## Data Cleaning

1. imdb data set: 52452 variables. Noticably, movie id, movie name, genre, runtime, rating, director, director id, star, star id, box office
2. Rotten Tomatoes Movie Reviews: 1444963 variables. movieid, review id, critic, is topcritic, originalscore, review text
3. Rotten Tomatoes Movie: 143258 variables. Movieid, audience score, tomato meter, runtime, genre, director, writer, box office

## Research Question 1

**How do specific movie attributes (genre, acting, cinematography, soundtrack, runtime, director, writer, star power) correlate with review scores and box office revenue? Which attributes are most important for each types of movie genre?**

Clean the data:

* Convert to **lowercase**
* Remove **punctuation, numbers, special characters**
* Remove **stopwords** (e.g., “the,” “is,” “and”)
* Use **lemmatization** (convert words to base form, e.g., “running” → “run”)

Prompt for chatgpt:

Analyze the following movie review and extract any key attributes related to the movie’s quality (e.g., acting, cinematography, soundtrack, pacing, effects, dialogue, or any other relevant aspects).  
Identify whether each attribute is mentioned positively or negatively.

Review: “The lead actor gave an unforgettable performance, but the camera work felt uninspired.”

Output: - Acting: Positive - Cinematography: Negative

**Use OpenAI for Attribute Extraction**

import openai import pandas as pd import time

API\_KEY = “your\_openai\_api\_key\_here”

# Function to analyze a review

def analyze\_review(review): prompt = f”“” Analyze the following movie review and extract any key attributes related to the movie’s quality (e.g., acting, cinematography, soundtrack, pacing, effects, dialogue, screenplay, editing, lighting, or any other relevant aspects). Identify whether each attribute is mentioned positively or negatively.

Review: "{review}"  
  
Output format:  
Attribute: Sentiment (Positive/Negative)  
  
Example:  
- Acting: Positive  
- Cinematography: Negative  
- Sound Design: Positive  
- Screenplay: Negative  
"""  
  
try:  
 response = openai.ChatCompletion.create(  
 model="gpt-4-turbo",  
 messages=[{"role": "system", "content": "You are an expert at analyzing movie reviews."},  
 {"role": "user", "content": prompt}],  
 temperature=0.2  
 )  
 return response["choices"][0]["message"]["content"]  
  
except Exception as e:  
 print("Error:", e)  
 return None

**Apply GPT to reviews using batching Code below:**

# Load dataset

df = pd.read\_csv(“your\_reviews.csv”)

# Clean the reviews (Convert to lowercase, remove special characters)

def clean\_text(text): import re text = text.lower() # Convert to lowercase text = re.sub(r”[^\w\s]“,”“, text) # Remove punctuation return text

df[“cleaned\_reviews”] = df[“review\_text”].apply(clean\_text)

**Process Reviews in Batches**

# Function to process reviews in batches

def process\_reviews\_in\_batches(reviews, batch\_size=5, delay=1): results = []

for i in range(0, len(reviews), batch\_size):  
 batch = reviews[i:i+batch\_size]  
 batch\_results = []  
   
 for review in batch:  
 result = analyze\_review(review)  
 batch\_results.append(result)  
 time.sleep(delay) # Prevent hitting API rate limits  
   
 results.extend(batch\_results)  
  
return results

# Apply function to all reviews

df[“attribute\_analysis”] = process\_reviews\_in\_batches(df[“cleaned\_reviews”].dropna().tolist())

Once I get this done I hope to analyze and group everything by genre

## **Research Question 2**

**How do certain actors, writers, directors, run times, and genres impact review score and box office performance?**

1. Conduct a correlation analysis on run time and reviews
2. Graph and list top 10 actors- visualize the data
3. Train a predictive model to predict box office revenue

# Select features

features <- c(“director\_experience”, “director\_avg\_review”, “actor\_experience”, “actor\_box\_office\_avg”, “normalized\_runtime”)

# Split data into training and testing sets (80/20 split)

set.seed(42) trainIndex <- createDataPartition(movies$box\_office, p = 0.8, list = FALSE) train <- movies[trainIndex,] test <- movies[-trainIndex,]

# Train linear regression model

model <- lm(box\_office ~ ., data = train %>% select(all\_of(features), box\_office))

# Model summary

summary(model)

# Predict on test set

predictions <- predict(model, newdata = test %>% select(all\_of(features)))

# Evaluate model performance

mse <- mean((test$box\_office - predictions)^2) print(paste(“Mean Squared Error:”, mse))

Find which attributes matter the most for box office and reviews

# Extract feature importance

importance <- as.data.frame(varImp(model)) importance$feature <- rownames(importance)

# Plot feature importance

ggplot(importance, aes(x = reorder(feature, Overall), y = Overall)) + geom\_bar(stat = “identity”, fill = “blue”) + coord\_flip() + labs(title = “Feature Importance in Box Office Prediction”, x = “Feature”, y = “Importance”)

## Research Question 3  
  
\*\*Do higher review scores lead to higher box office revenue?\*\*  
  
1. Data cleaning (no missing values)  
2. Calculate the correlation coefficent  
3. regression analysis  
  
  
# Build linear regression model  
model <- lm(box\_office ~ review\_score, data = movies)  
  
# Model summary  
summary(model)

1. Check to see if there are other things that may be contributing to the score other than reviews.

# Multiple regression model

advanced\_model <- lm(box\_office ~ review\_score + normalized\_runtime + director\_experience + genre\_Action + genre\_Comedy, data = movies)

# Model summary

summary(advanced\_model)

```

### Research Question 4

**Predicting Box Office Revenue Using Review Sentiment and Movie Attributes**

#### Model Concept:

You could build a **machine learning model** to predict **box office revenue** based on features like **review sentiment (critic vs. audience), genre, director, writer, runtime, star power**, and even **movie review counts**. The idea is to train the model using historical data to predict box office revenue for new or upcoming movies, providing valuable insights for studios or production companies.

#### Steps to Build the Model:

**a. Data Preprocessing:**

* **Sentiment Analysis on Reviews**: Use NLP techniques (like **VADER** or **BERT** for sentiment analysis) to assign sentiment scores to each review in your dataset (critic and audience reviews). This will give you numerical features like **positive**, **neutral**, or **negative** sentiment scores.
* **Feature Engineering**: Extract and create meaningful features from the raw data (e.g., **genre** (one-hot encode), **director/actor star power** (based on previous box office performance or social media presence), **runtime** (continuous feature)).
* **Data Cleaning**: Handle missing values, outliers, and irrelevant variables (e.g., duplicate reviews).

**b. Model Selection:**

* **Linear Regression**: A good starting point, as it can help you understand how each feature (e.g., sentiment score, director, genre) impacts box office revenue.
* **Random Forest or Gradient Boosting**: These tree-based algorithms can handle non-linear relationships and are often better at capturing the complexity of real-world data.
* **Neural Networks**: If you want to use deep learning, you could train a neural network to capture more complex patterns, especially if you include unstructured data like review text (processed via NLP).

**c. Training the Model:**

* **Split the data** into training and testing sets (e.g., 80% for training, 20% for testing).
* **Train your model** on the training data. For machine learning, you might need to tune hyperparameters (e.g., using **GridSearchCV** for Random Forests or Gradient Boosting) to improve accuracy.

**d. Model Evaluation:**

* **Performance Metrics**: Use metrics like **Mean Absolute Error (MAE)** or **Root Mean Squared Error (RMSE)** to evaluate the model’s accuracy in predicting box office revenue.
* **Cross-validation**: Apply **cross-validation** (k-fold) to avoid overfitting and ensure the model generalizes well to unseen data.

### Other Review Questions if there is time

1. How does the sentiment of critic and audience reviews compare, and does it correlate with box office revenue?
2. impact of the reviewer’s status” (e.g., top critics vs. general critics) on box office performance

# **Why This Study Matters**

### **Better Prediction Models for Studios and Marketers**

Studios can use my findings to optimize production and marketing budgets, maximizing return on investment.

### **Understanding What Truly Drives Box Office Success**

My study helps differentiate between factors that drive critical acclaim vs. box office revenue.  
If review scores don’t always correlate with financial success, studios can make more informed decisions about balancing artistic quality with commercial appeal.

### **Advancing Machine Learning Applications in Entertainment**

Using AI for sentiment analysis and predictive modeling isn’t just about movies; it has applications in predicting product success across industries (e.g., books, music, video games).  
My approach, especially with NLP techniques, could lead to more refined models in entertainment analytics.

### **Reassessing the Role of Critics and Audience Reviews**

If top critics’ opinions don’t align with audience sentiment or revenue outcomes, it raises questions about their influence.  
My study could provide insights into whether studios should focus more on audience feedback rather than critic scores.

### **Contributing to Film Studies and Storytelling Science**

Identifying which attributes resonate most within each genre can help filmmakers refine storytelling, cinematography, and character development strategies.  
It can help film schools and aspiring filmmakers understand how different elements contribute to both artistic and financial success.

### **Has This All Been Done Before?**

While there have been past studies on box office prediction, my work expands on existing research by integrating a broader range of variables, using advanced sentiment analysis, and testing multiple predictive models.  
The industry constantly evolves (e.g., streaming vs. theatrical releases, changing audience preferences), making continuous research necessary. Additionally many of these studies were done before AI and so I have a chance to broaden this field as a whole.

### Limitations

While there are many limitations I believe that the biggest is that my data doesn’t have the budget for these different movies. I am unfortunately assuming that a bigger box office is better. However, a movie may have a bigger box office than another but because of a more expensive budget it could be less successful. This is a huge drawback to my data.