1. Introduction

This document serves as the Data Appendix for the predictive modeling project "Sentiment Analysis of Pre-Released Critic Reviews for Box Office Prediction." It provides a structured overview of the datasets and transformations applied, from raw input data to final model performance tracking.

2. Data Pipeline Workflow

Step 1: Input Data Files

IMDb Reviews Dataset (imdb_reviews.csv)

- Unit of Observation: Each row represents a single movie review.
- Key Variables:
  - Movie_id – Unique IMDb identifier
  - Review_text – Full text of the review
- Purpose: Contains critic reviews that we process for sentiment analysis.
- Processing Steps: Scraped IMDb critic reviews using web_scraper.py

Hand-Picked Movie Data (hand_picked_movie_data.csv)

- Unit of Observation: Each row represents a selected movie.
- Key Variables:
  - Rank - Rank of the movie based on opening weekend earnings
  - Release - Movie title
  - Opening - Opening weekend revenue (USD)
  - Open - Release month
  - Distributor - Film distributor/studio
  - Movie_id - Unique IMDb identifier
- Purpose: A curated set of movies to cross-check aggregated sentiment data.
- Processing Steps: Manually collected data from the Domestic Box Office for 2024 movies on Box Office Mojo.

Step 2: Intermediate Data Processing

Sentiment Results (sentiment_results.csv)

- Unit of Observation: Each row represents a sentiment analysis result for a single review.

- Key Variables:
    - movie_id – Unique IMDb identifier for the movie
    - review_text – Full text of the critic’s review
    - neg – Negative sentiment score (fraction of the review expressing negativity, 0 to 1)
    - neu – Neutral sentiment score (fraction of the review expressing neutrality, 0 to 1)
    - pos – Positive sentiment score (fraction of the review expressing positivity, 0 to 1)
    - compound – Overall sentiment score (-1 to +1) computed by VADER
- Processing Steps:
    - Applied VADER Sentiment Analysis using sentimentAnalyzer.py
    - Extracted sentiment scores (Negative, Neutral, Positive, Compound)

Review Dataset (review_dataset.csv)

- Unit of Observation: Each row represents a single critic review.
- Key Variables:
    - Movie_id – Unique IMDb identifier for the movie
    - Review_text – Full text of the critic’s review
    - Negative_sentiment_score – Computed negative sentiment score (VADER)
    - Neutral_sentiment_score – Computed neutral sentiment score (VADER)
    - Positive_sentiment_score – Computed positive sentiment score (VADER)
    - Compound_sentiment_score – Computed compound sentiment score (-1 to +1)
    - Rank – Rank of the movie based on opening weekend earnings
    - Name – Movie title
    - Opening_earnings – Opening weekend revenue in USD
    - Release_month – Month in which the movie was released
    - Distributor – The studio or company responsible for distributing the movie
- Processing Steps:
    - Extracted structured sentiment scores.
    - Mapped reviews to movies using Movie_id.

Aggregated Movie Reviews (aggregated_movie_reviews.csv)

- Unit of Observation: Each row represents a movie with aggregated sentiment scores.
- Key Variables:
    - Movie_id – Unique IMDb identifier
    - Negative_sentiment_score – Average negative sentiment score across all critic reviews for the movie
    - Neutral_sentiment_score – Average neutral sentiment score across all critic reviews for the movie

    - Positive_sentiment_score – Average positive sentiment score across all critic reviews for the movie
    - Compound_sentiment_score – Average compound sentiment score (-1 to +1) across all critic reviews
    - Rank – Rank of the movie based on opening weekend earnings
    - Name – Movie title
    - Opening_earnings – Opening weekend revenue in USD
    - Release_month – Month in which the movie was released
    - Distributor – The studio or company responsible for distributing the movie
- Processing Steps:
    - Grouped sentiment results by Movie_id
    - Computed average sentiment scores per movie
    - Merged metadata (Distributor, Release Month)

Step 3: Cleaned Analysis Data

Cleaned Movie Data (cleaned_movie_data.csv)

- Unit of Observation: Each row represents a single movie with all final variables.
- Key Variables:
    - Negative_sentiment_score – Average negative sentiment score (VADER) across all critic reviews for the movie
    - Neutral_sentiment_score – Average neutral sentiment score (VADER) across all critic reviews
    - Positive_sentiment_score – Average positive sentiment score (VADER) across all critic reviews
    - Compound_sentiment_score – Overall sentiment score (-1 to +1, aggregated)
    - Rank – Movie's ranking based on opening weekend earnings
    - Opening_earnings – Opening weekend box office revenue (USD)
    - Release_month_* – One-hot encoded variables indicating the movie's release month (e.g., Release_month_Jan, Release_month_Jul)
    - Distributor_* – One-hot encoded variables indicating the movie's distributor (e.g., Distributor_Walt Disney Studios Motion Pictures, Distributor_Warner Bros.)
- Processing Steps:
    - Removed missing or inconsistent data to ensure a clean dataset.
    - Performed feature engineering by applying one-hot encoding to categorical variables (Release_month, Distributor).
    - Standardized numerical features to normalize values for better model performance.
    - Ensured final dataset alignment with previous processing steps for compatibility in model training.

Step 4: Model Performance Tracking

Model Performance Results (model_performance_tracking.csv)

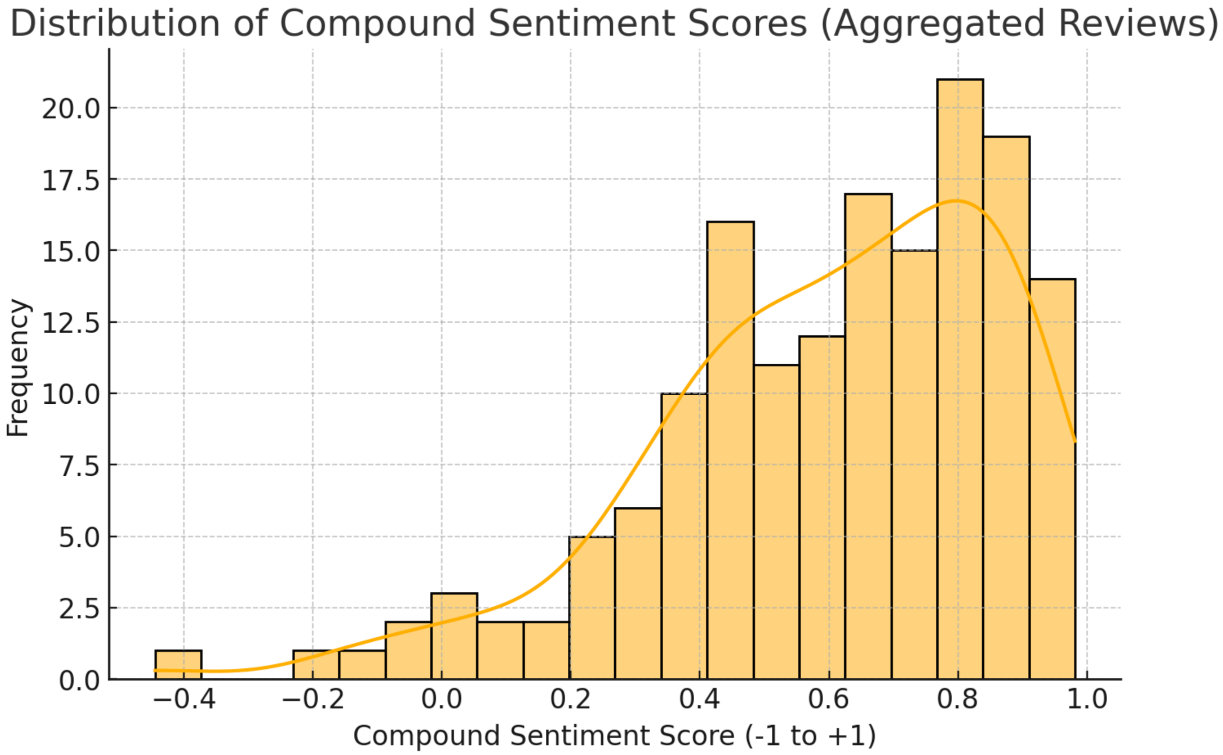
- Unit of Observation: Each row represents a trained model and its evaluation metrics.
- Key Variables:
  1. MAE – Mean Absolute Error, measuring the average absolute difference between predicted and actual values (lower is better).
  2. MSE – Mean Squared Error, measuring the squared differences between predicted and actual values (lower is better).
  3. R2 Score – Coefficient of determination, indicating how well the model explains variance in earnings (higher is better).
- Processing Steps:
  1. Trained multiple regression models (e.g., Linear Regression, Random Forest, XGBoost) on cleaned_movie_data.csv.
  2. Evaluated model performance using MAE, MSE, and R² Score.

3. Summary Statistics and Visualizations

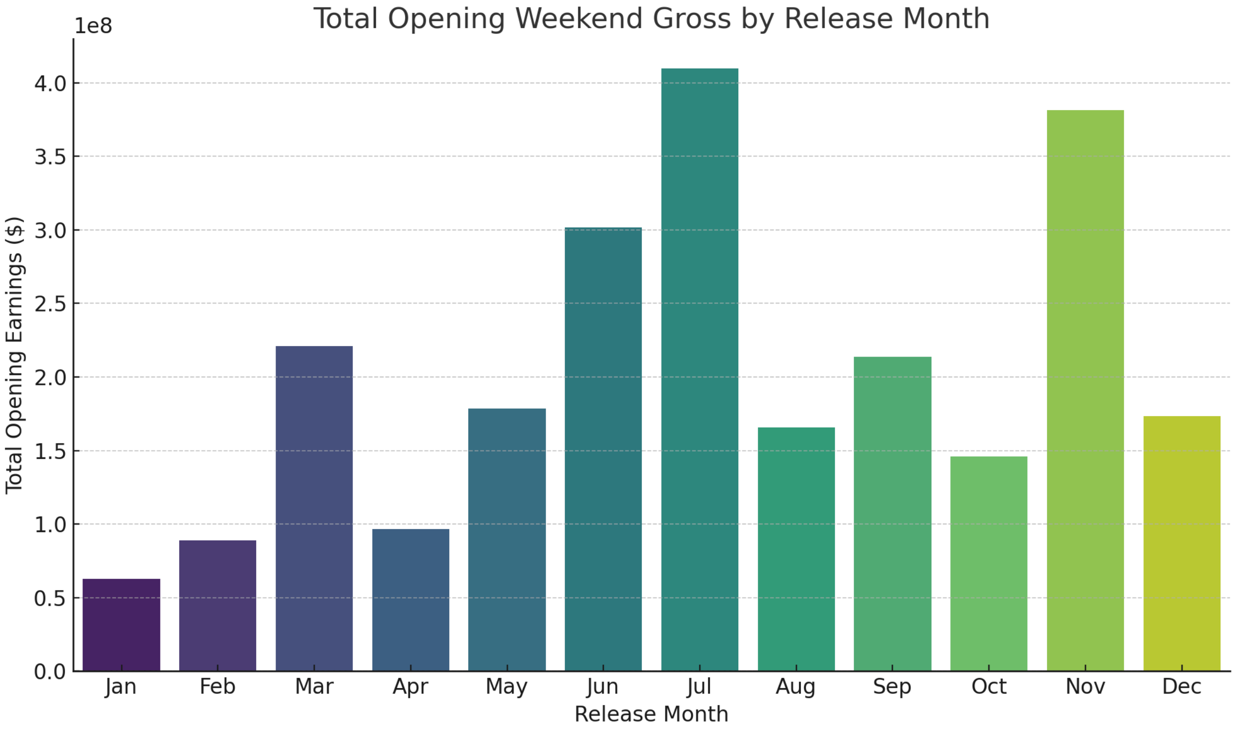
Cleaned Movie Data Summary:

	negative_sentimet	neutral_sentiment	positive_sentiment	compound_sentiment	rank	opening_earnings
count	158.0	158.0	158.0	158.0	158.0	158.0
mean	0.07754216560509554	0.7489391082802548	0.17351668789808916	0.6012429044585987	88.73885350318471	15535271.751592357
std	0.024769828377366566	0.024819485970001946	0.030054979692960625	0.26817737485921855	55.37172129516961	29824871.417857666
min	0.03188	0.70384	0.10668	-0.444384	1.0	18259.0
max	0.15596	0.81472	0.2602	0.981236	200.0	211435291.0

Histogram of Compound Sentiment Scores from the Aggregated Movie Reviews Dataset:



Box plot of Opening Weekend Gross by Release Month


Total Opening Weekend Gross by Release Month
1e8
Total Opening Earnings ($)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Release Month