Contents

Deliverable

This document provides a data exploration report for iMBD datasets. It delivers the following outputs:

1. A comprehensive overview of the raw data, enabling users to grasp the dataset's structure, content, and variable descriptions.
2. A publication-ready document presented in a well-formatted layout, including explanatory text, tables, and visualizations.

The main goal of this report is to promote transparency and reproducibility by helping readers understand the cleaned dataset that forms the basis for the subsequent regression analysis.

Data Overview

This project uses data from the IMDb (Internet Movie Database), a publicly available source containing information on movies, including metadata such as titles, release years, genres, runtimes, and ratings.

The data were originally obtained from IMDb's open dataset and subsequently processed within our workflow for reproducibility and consistency. The key dataset used for analysis is the cleaned version of the IMDb movie data, stored as movies_final_clean.csv.

The following data files are utilized in this project:

1. model_1.csv – Contains the regression model results and intermediate outputs used for prediction and visualization.
2. movies_final_clean.csv – Includes the cleaned and processed movie data, with variables such as title, release_year, runtime_minutes, and average_rating.

All files are generated through automated R scripts as part of the project's Makefile pipeline, ensuring that data preparation and analysis steps can be reproduced seamlessly.

Only movies with complete information on runtime, release year, and IMDb rating were retained in the final dataset to ensure valid and interpretable regression results.

Data Preparation Summary

The data preparation process was conducted in R, using packages such as dplyr, tidyr, and ggplot2, and follows a structured, reproducible workflow designed to produce the cleaned dataset movies_final_clean.csv.

The main steps include:

1. Data Loading and Inspection

Raw IMDb data files were imported into R and inspected for completeness, consistency, and correct data types. Missing or invalid entries (e.g., missing runtimes or ratings) were identified and addressed.

2. Variable Selection

From the raw dataset, only variables relevant to the regression analysis were retained — including: title, release_year, runtime_minutes, and average_rating. Additional variables such as genre or votes were excluded from this stage to focus the analysis on runtime and rating relationships.

3. Data Cleaning

Entries with missing or implausible values (e.g., runtime_minutes < 10 or > 400) were removed.

Ratings were checked to ensure they fall within the valid IMDb scale (1.0–10.0).

Movie titles were standardized, and duplicate entries were dropped.

4. Feature Preparation

To enable year-adjusted comparisons, the variable release_year was retained as a control variable for regression modeling. This allows for an examination of the relationship between movie length and IMDb rating while controlling for differences across years.

5. Output Generation

The cleaned dataset was exported as movies_final_clean.csv, forming the foundation for all subsequent statistical analyses and visualizations. Intermediate model outputs, including predictions and confidence intervals, were stored in model_1.csv for later use in graphical presentation.

Variable Description

Variable	Description	Data Type
`title`	Official movie title as listed on IMDb	Character
`release_year`	Year in which the movie was released	Numeric
`runtime_minutes`	Total runtime of the movie in minutes	Numeric
`average_rating`	IMDb user rating on a 1–10 scale	Numeric
`num_votes`	Number of user votes contributing to the average rating	Numeric
`genre`	Primary genre classification from IMDb data	Character

This structured dataset provides the foundation for modeling and visualization, enabling the examination of how movie runtime relates to audience ratings, while accounting for variation in release year.

Summary Statistics

Before running any statistical models, exploratory analysis was performed to understand the main characteristics of the cleaned IMDb dataset.

- The dataset includes thousands of movies released over multiple decades.
- The average IMDb rating is around 6.5, with most values falling between 5.5 and 7.5.
- The average runtime is approximately 100 minutes, with most movies ranging between 80 and 120 minutes.

- There is noticeable variability in both runtime and rating, providing sufficient variation for regression analysis.
- Outliers (e.g., extremely short or very long films) were inspected but retained if verified as valid entries.
- These descriptive findings suggest a generally balanced dataset, suitable for modeling the relationship between runtime and IMDb rating.

Key Insights

1. The cleaned IMDb dataset provides a robust foundation for statistical analysis, containing a large number of films with complete information on runtime, release year, and average rating.
2. Descriptive statistics and visualizations suggest that movie runtimes are generally concentrated between 80 and 120 minutes, while ratings cluster around the mid-to-high range of the IMDb scale (5.5–7.5).
3. Preliminary visual exploration reveals a modest positive relationship between runtime and IMDb rating — indicating that longer films may receive slightly higher audience ratings on average.
4. Controlling for release year is likely important, as the relationship between runtime and rating may vary across decades and film industry trends.
5. The dataset shows sufficient variability and data quality to support regression-based modeling, providing a sound basis for testing hypotheses about runtime and rating dynamics.

Regression Framework and Future Analysis

Following the exploratory analysis, the next stage of the project focuses on formally testing the relationship between movie runtime and IMDb rating, while accounting for release year as a control variable.

The primary research question guiding this stage is:

"How does a movie's runtime influence its IMDb rating, and does this relationship hold after adjusting for release year?"

To address this question, a Multiple Linear Regression (MLR) model is estimated using the cleaned dataset movies_final_clean.csv. The model quantifies how changes in runtime are associated with differences in predicted IMDb ratings, while holding release year constant.

Model diagnostics and visualization outputs are generated using packages such as modelsummary, broom, and ggpredict. Predicted relationships and confidence intervals are visualized through figures such as "Predicted IMDb Rating vs Runtime", stored in the output directory for reproducibility.

All model outputs and graphical summaries are compiled into the final report, ensuring a transparent and replicable analytical workflow.