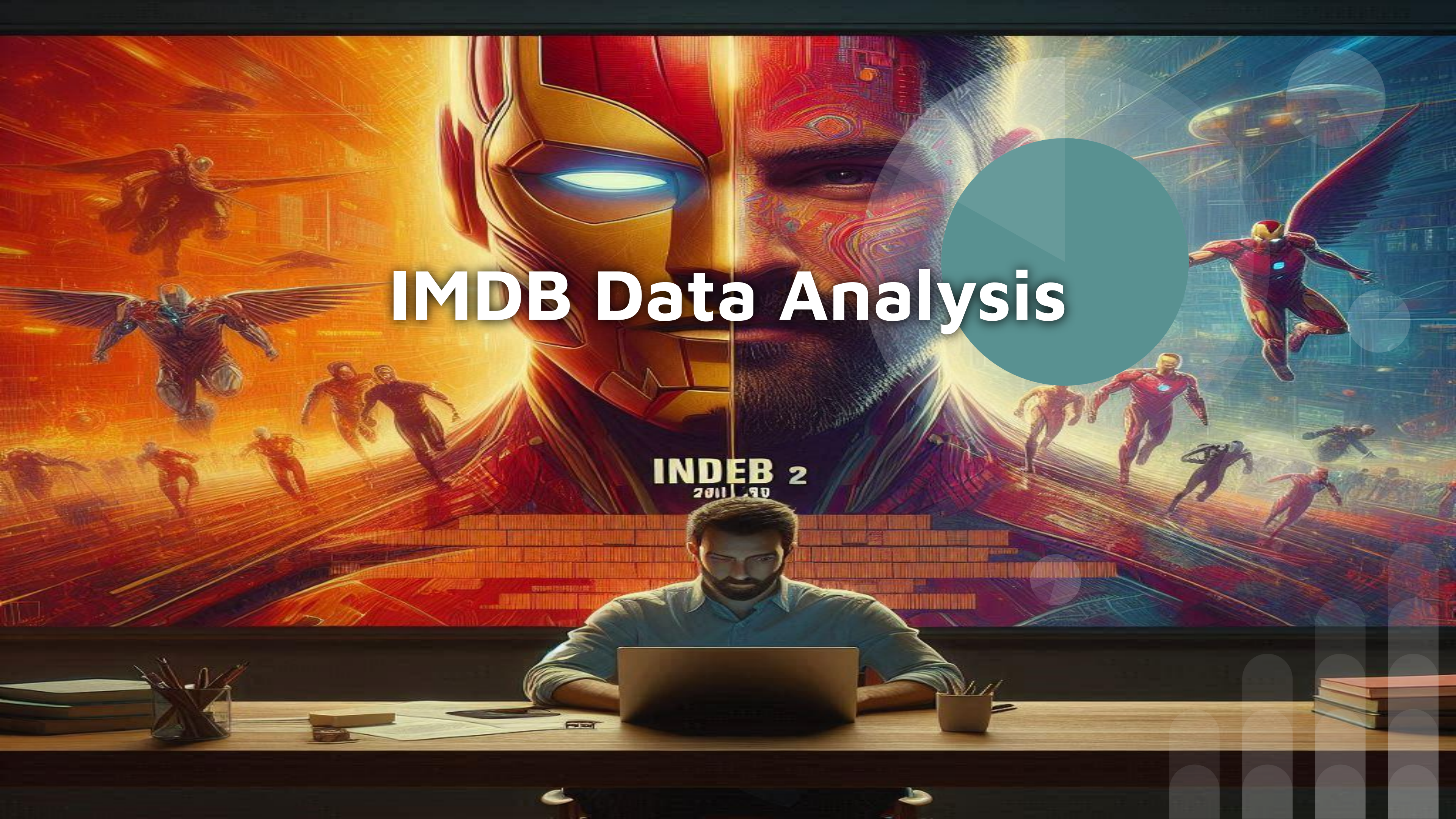
# IMDB Data Analysis



# Understanding the project

### Phase 0 : Gathering data

About the dataset and preprocessing challenges

### Phase 1 : Analyzing Data

Analyze and visualize data to understand the key characteristics, uncover patterns

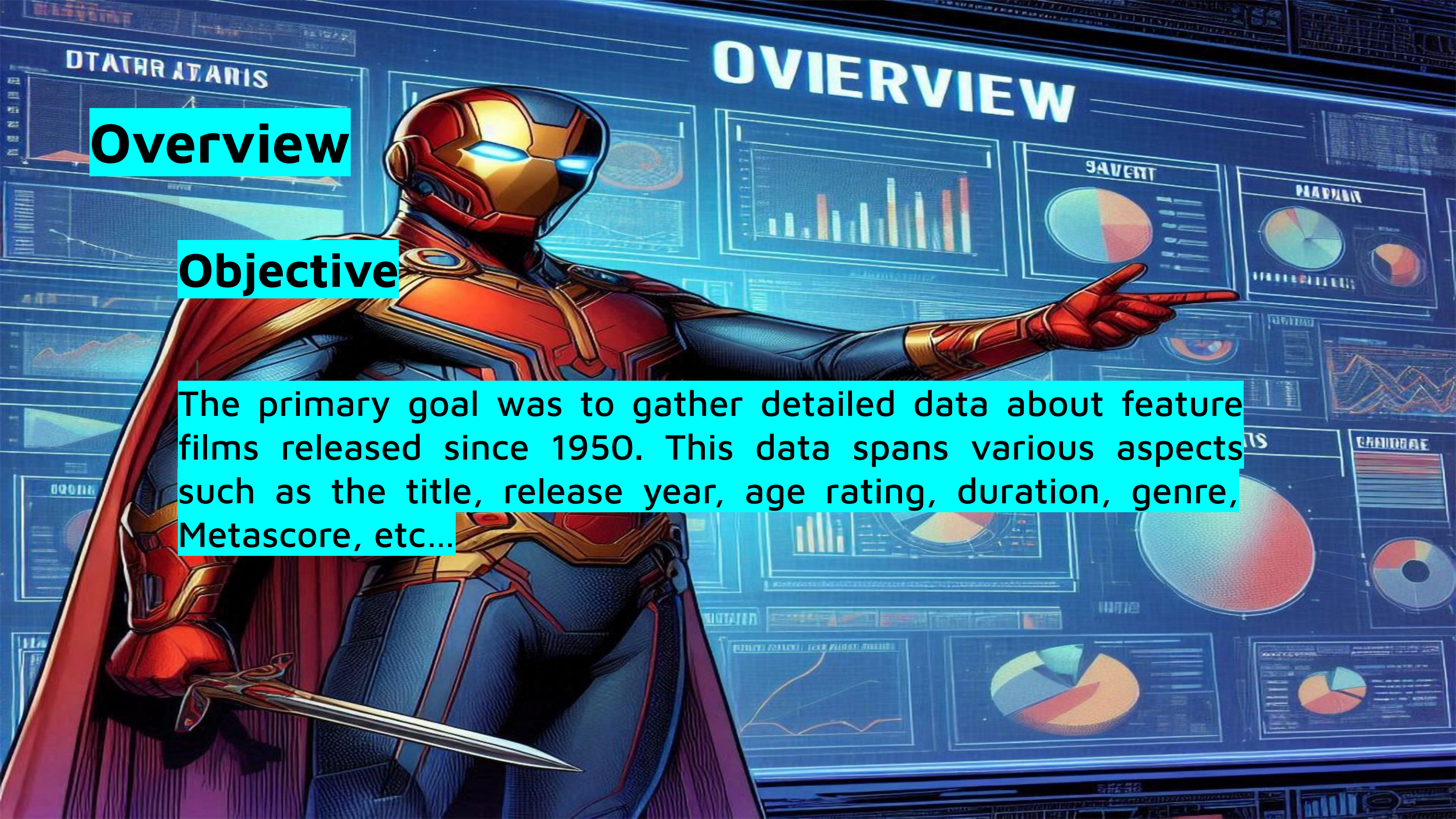
### Phase 2: Modeling

Predicting rating, profits and Age Restriction.

# Overview

## Objective

The primary goal was to gather detailed data about feature films released since 1950. This data spans various aspects such as the title, release year, age rating, duration, genre, Metascore, etc...

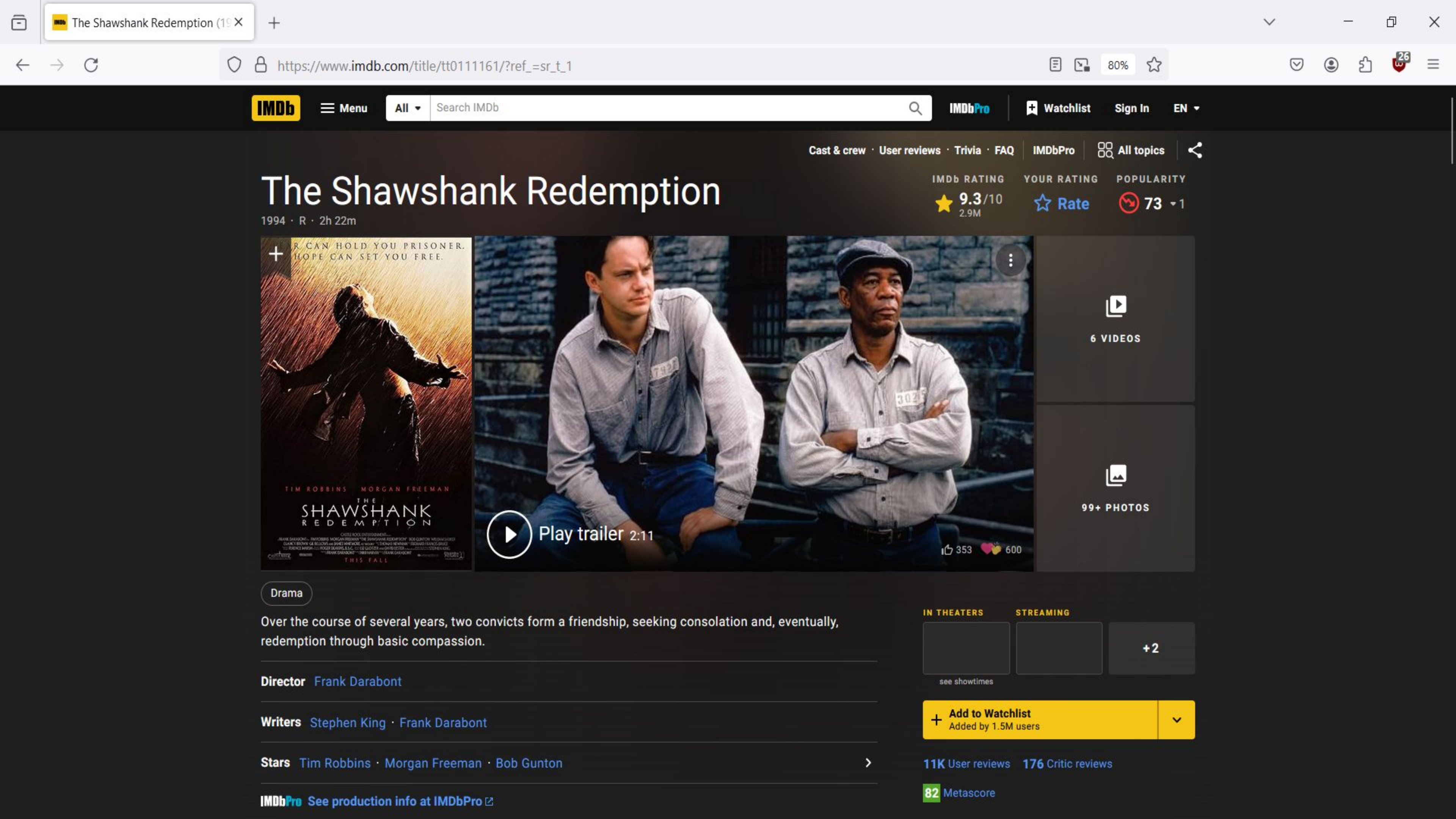


# About Dataset and Gathering

# The way we gathered data

To automate the browser and interact with the web pages, we used Selenium.

Beautiful Soup Helped us parse the HTML and extract the required data.



# The Shawshank Redemption

1994 · R · 2h 22m

Drama

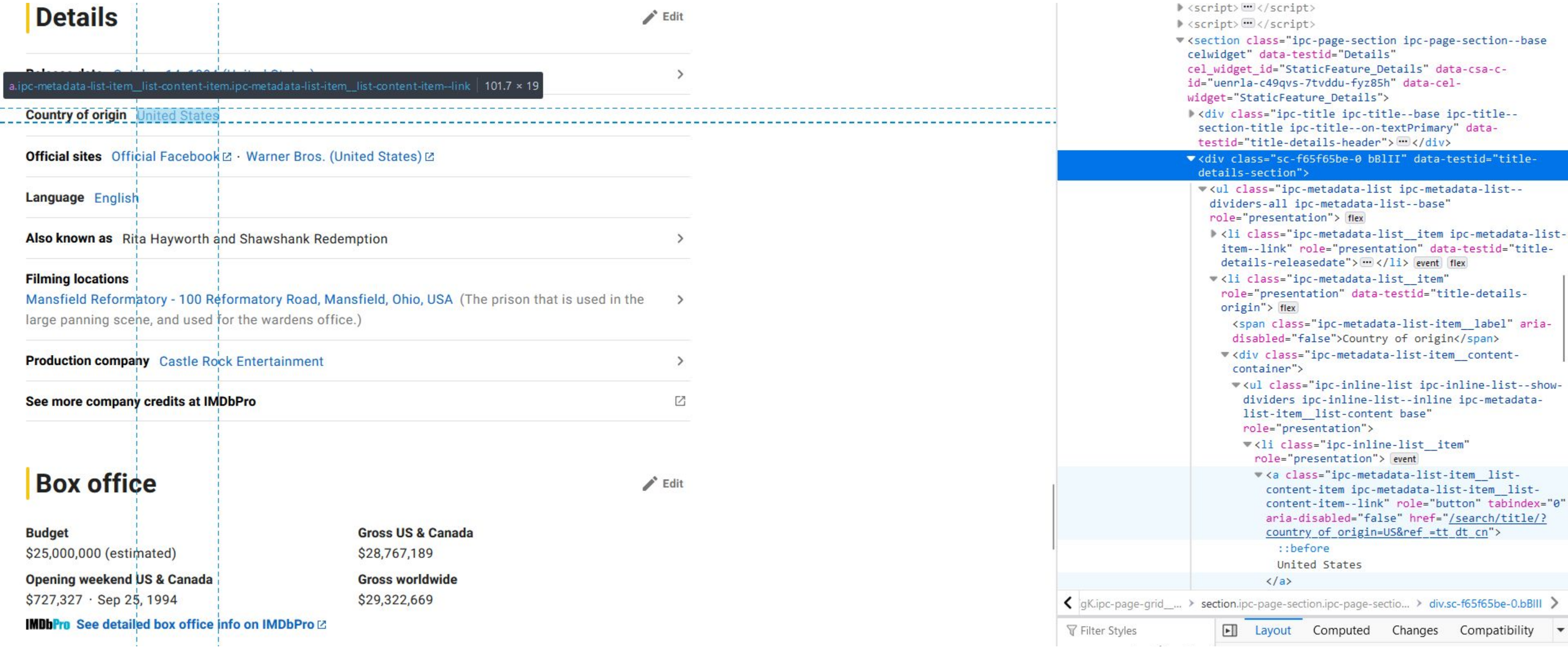
Over the course of several years, two convicts form a friendship, seeking consolation and, eventually, redemption through basic compassion.

**Director** Frank Darabont

**Writers** Stephen King · Frank Darabont

**Stars** Tim Robbins · Morgan Freeman · Bob Gunton

IMDbPro See production info at IMDbPro



## Details

Edit

Release date October 14, 1994 (United States)

a.ipc-metadata-list-item__list-content-item.ipc-metadata-list-item__list-content-item--link | 101.7 × 19

**Country of origin** United States

**Official sites** Official Facebook · Warner Bros. (United States)

**Language** English

**Also known as** Rita Hayworth and Shawshank Redemption

**Filming locations**
Mansfield Reformatory - 100 Reformatory Road, Mansfield, Ohio, USA (The prison that is used in the large panning scene, and used for the wardens office.)

**Production company** Castle Rock Entertainment

**See more company credits at IMDbPro**

## Box office

Edit

**Budget**
$25,000,000 (estimated)

**Gross US & Canada**
$28,767,189

**Opening weekend US & Canada**
$727,327 · Sep 25, 1994

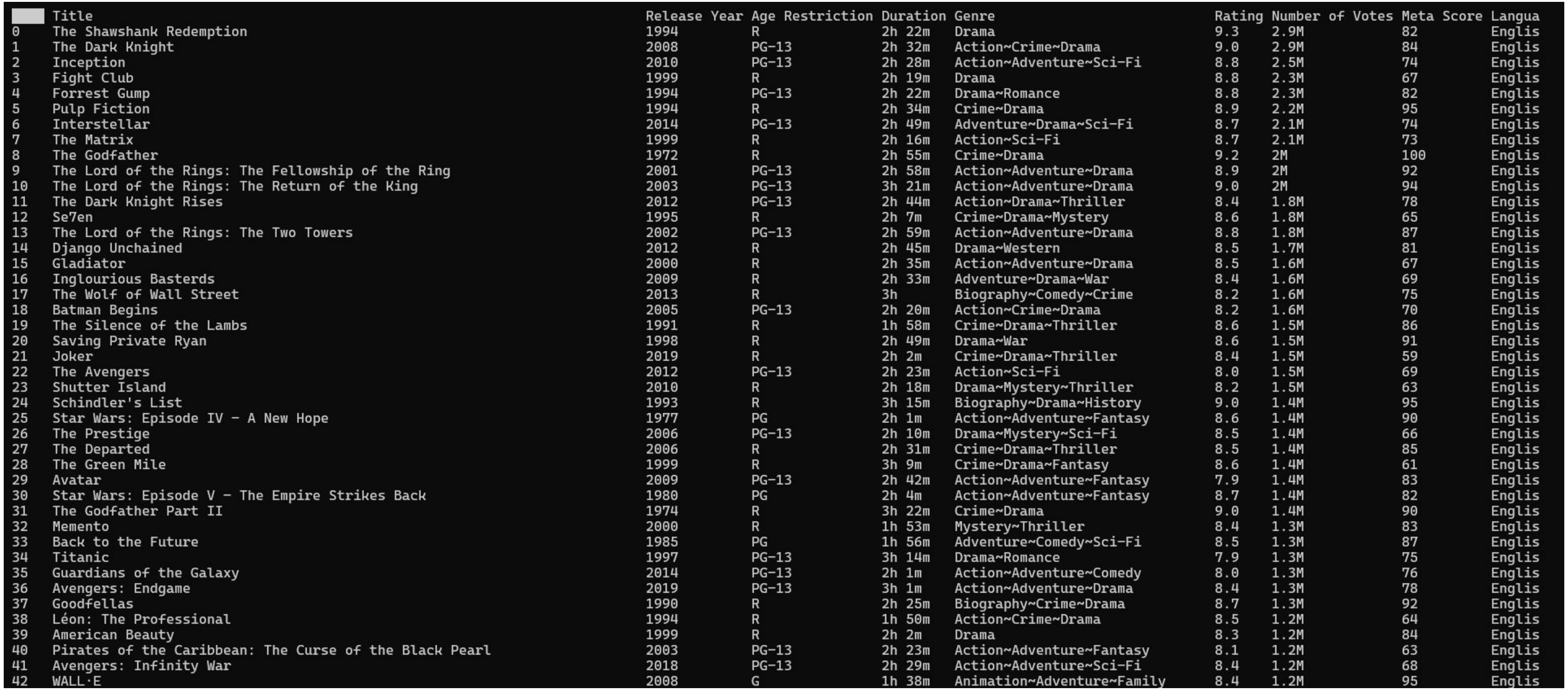
**Gross worldwide**
$29,322,669

IMDbPro **See detailed box office info on IMDbPro**

```
▶ <script> ⋯ </script>
▶ <script> ⋯ </script>
▼ <section class="ipc-page-section ipc-page-section--base
  celwidget" data-testid="Details"
  cel_widget_id="StaticFeature_Details" data-csa-c-
  id="uenrla-c49qvs-7tvddu-fyz85h" data-cel-
  widget="StaticFeature_Details">
  ▶ <div class="ipc-title ipc-title--base ipc-title--
    section-title ipc-title--on-textPrimary" data-
    testid="title-details-header"> ⋯ </div>
  ▼ <div class="sc-f65f65be-0 bBlII" data-testid="title-
    details-section">
    ▼ <ul class="ipc-metadata-list ipc-metadata-list--
      dividers-all ipc-metadata-list--base"
      role="presentation"> flex
      ▶ <li class="ipc-metadata-list__item ipc-metadata-list-
        item--link" role="presentation" data-testid="title-
        details-releasedate"> ⋯ </li> event flex
      ▼ <li class="ipc-metadata-list__item"
        role="presentation" data-testid="title-details-
        origin"> flex
          <span class="ipc-metadata-list-item__label" aria-
          disabled="false">Country of origin</span>
        ▼ <div class="ipc-metadata-list-item__content-
          container">
          ▼ <ul class="ipc-inline-list ipc-inline-list--show-
            dividers ipc-inline-list--inline ipc-metadata-
            list-item__list-content base"
            role="presentation">
            ▼ <li class="ipc-inline-list__item"
              role="presentation"> event
              ▼ <a class="ipc-metadata-list-item__list-
                content-item ipc-metadata-list-item__list-
                content-item--link" role="button" tabindex="0"
                aria-disabled="false" href="/search/title/?
                country_of_origin=US&ref_=tt_dt_cn">
                  ::before
                  United States
                </a>
```

gK.ipc-page-grid__... › section.ipc-page-section.ipc-page-sectio... › div.sc-f65f65be-0.bBlII
Filter Styles
Layout
Computed
Changes
Compatibility

# Final Dataset (Only a part of it)



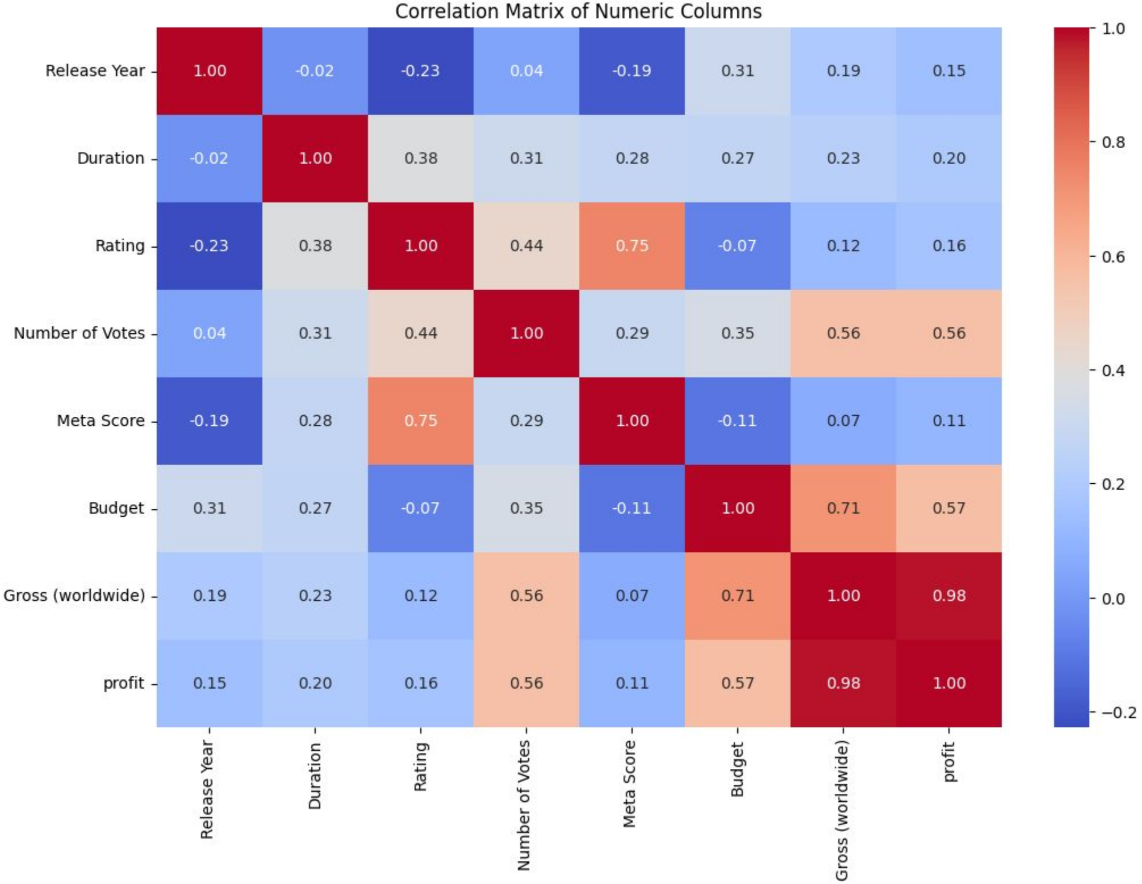
```
     Title                                                                    Release Year Age Restriction Duration Genre                           Rating Number of Votes Meta Score Langua
0    The Shawshank Redemption                                                 1994         R               2h 22m   Drama                           9.3    2.9M            82         Englis
1    The Dark Knight                                                          2008         PG-13           2h 32m   Action~Crime~Drama              9.0    2.9M            84         Englis
2    Inception                                                                2010         PG-13           2h 28m   Action~Adventure~Sci-Fi         8.8    2.5M            74         Englis
3    Fight Club                                                               1999         R               2h 19m   Drama                           8.8    2.3M            67         Englis
4    Forrest Gump                                                             1994         PG-13           2h 22m   Drama~Romance                   8.8    2.3M            82         Englis
5    Pulp Fiction                                                             1994         R               2h 34m   Crime~Drama                     8.9    2.2M            95         Englis
6    Interstellar                                                             2014         PG-13           2h 49m   Adventure~Drama~Sci-Fi          8.7    2.1M            74         Englis
7    The Matrix                                                               1999         R               2h 16m   Action~Sci-Fi                   8.7    2.1M            73         Englis
8    The Godfather                                                            1972         R               2h 55m   Crime~Drama                     9.2    2M              100        Englis
9    The Lord of the Rings: The Fellowship of the Ring                        2001         PG-13           2h 58m   Action~Adventure~Drama          8.9    2M              92         Englis
10   The Lord of the Rings: The Return of the King                            2003         PG-13           3h 21m   Action~Adventure~Drama          9.0    2M              94         Englis
11   The Dark Knight Rises                                                    2012         PG-13           2h 44m   Action~Drama~Thriller           8.4    1.8M            78         Englis
12   Se7en                                                                    1995         R               2h 7m    Crime~Drama~Mystery             8.6    1.8M            65         Englis
13   The Lord of the Rings: The Two Towers                                    2002         PG-13           2h 59m   Action~Adventure~Drama          8.8    1.8M            87         Englis
14   Django Unchained                                                         2012         R               2h 45m   Drama~Western                   8.5    1.7M            81         Englis
15   Gladiator                                                                2000         R               2h 35m   Action~Adventure~Drama          8.5    1.6M            67         Englis
16   Inglourious Basterds                                                     2009         R               2h 33m   Adventure~Drama~War             8.4    1.6M            69         Englis
17   The Wolf of Wall Street                                                  2013         R               3h       Biography~Comedy~Crime          8.2    1.6M            75         Englis
18   Batman Begins                                                            2005         PG-13           2h 20m   Action~Crime~Drama              8.2    1.6M            70         Englis
19   The Silence of the Lambs                                                 1991         R               1h 58m   Crime~Drama~Thriller            8.6    1.5M            86         Englis
20   Saving Private Ryan                                                      1998         R               2h 49m   Drama~War                       8.6    1.5M            91         Englis
21   Joker                                                                    2019         R               2h 2m    Crime~Drama~Thriller            8.4    1.5M            59         Englis
22   The Avengers                                                             2012         PG-13           2h 23m   Action~Sci-Fi                   8.0    1.5M            69         Englis
23   Shutter Island                                                           2010         R               2h 18m   Drama~Mystery~Thriller          8.2    1.5M            63         Englis
24   Schindler's List                                                         1993         R               3h 15m   Biography~Drama~History         9.0    1.4M            95         Englis
25   Star Wars: Episode IV - A New Hope                                       1977         PG              2h 1m    Action~Adventure~Fantasy        8.6    1.4M            90         Englis
26   The Prestige                                                             2006         PG-13           2h 10m   Drama~Mystery~Sci-Fi            8.5    1.4M            66         Englis
27   The Departed                                                             2006         R               2h 31m   Crime~Drama~Thriller            8.5    1.4M            85         Englis
28   The Green Mile                                                           1999         R               3h 9m    Crime~Drama~Fantasy             8.6    1.4M            61         Englis
29   Avatar                                                                   2009         PG-13           2h 42m   Action~Adventure~Fantasy        7.9    1.4M            83         Englis
30   Star Wars: Episode V - The Empire Strikes Back                           1980         PG              2h 4m    Action~Adventure~Fantasy        8.7    1.4M            82         Englis
31   The Godfather Part II                                                    1974         R               3h 22m   Crime~Drama                     9.0    1.4M            90         Englis
32   Memento                                                                  2000         R               1h 53m   Mystery~Thriller                8.4    1.3M            83         Englis
33   Back to the Future                                                       1985         PG              1h 56m   Adventure~Comedy~Sci-Fi         8.5    1.3M            87         Englis
34   Titanic                                                                  1997         PG-13           3h 14m   Drama~Romance                   7.9    1.3M            75         Englis
35   Guardians of the Galaxy                                                  2014         PG-13           2h 1m    Action~Adventure~Comedy         8.0    1.3M            76         Englis
36   Avengers: Endgame                                                        2019         PG-13           3h 1m    Action~Adventure~Drama          8.4    1.3M            78         Englis
37   Goodfellas                                                               1990         R               2h 25m   Biography~Crime~Drama           8.7    1.3M            92         Englis
38   Léon: The Professional                                                   1994         R               1h 50m   Action~Crime~Drama              8.5    1.2M            64         Englis
39   American Beauty                                                          1999         R               2h 2m    Drama                           8.3    1.2M            84         Englis
40   Pirates of the Caribbean: The Curse of the Black Pearl                   2003         PG-13           2h 23m   Action~Adventure~Fantasy        8.1    1.2M            63         Englis
41   Avengers: Infinity War                                                   2018         PG-13           2h 29m   Action~Adventure~Sci-Fi         8.4    1.2M            68         Englis
42   WALL·E                                                                   2008         G               1h 38m   Animation~Adventure~Family      8.4    1.2M            95         Englis
```

# Exploratory Data Analysis

# Correlation

## Observations

- The profit , gross , budget and number of votes are highly dependable.
- The Rating and Metascore are highly dependable.



Correlation Matrix of Numeric Columns

| | Release Year | Duration | Rating | Number of Votes | Meta Score | Budget | Gross (worldwide) | profit |
|---|---|---|---|---|---|---|---|---|
| Release Year | 1.00 | -0.02 | -0.23 | 0.04 | -0.19 | 0.31 | 0.19 | 0.15 |
| Duration | -0.02 | 1.00 | 0.38 | 0.31 | 0.28 | 0.27 | 0.23 | 0.20 |
| Rating | -0.23 | 0.38 | 1.00 | 0.44 | 0.75 | -0.07 | 0.12 | 0.16 |
| Number of Votes | 0.04 | 0.31 | 0.44 | 1.00 | 0.29 | 0.35 | 0.56 | 0.56 |
| Meta Score | -0.19 | 0.28 | 0.75 | 0.29 | 1.00 | -0.11 | 0.07 | 0.11 |
| Budget | 0.31 | 0.27 | -0.07 | 0.35 | -0.11 | 1.00 | 0.71 | 0.57 |
| Gross (worldwide) | 0.19 | 0.23 | 0.12 | 0.56 | 0.07 | 0.71 | 1.00 | 0.98 |
| profit | 0.15 | 0.20 | 0.16 | 0.56 | 0.11 | 0.57 | 0.98 | 1.00 |

# Understanding the market

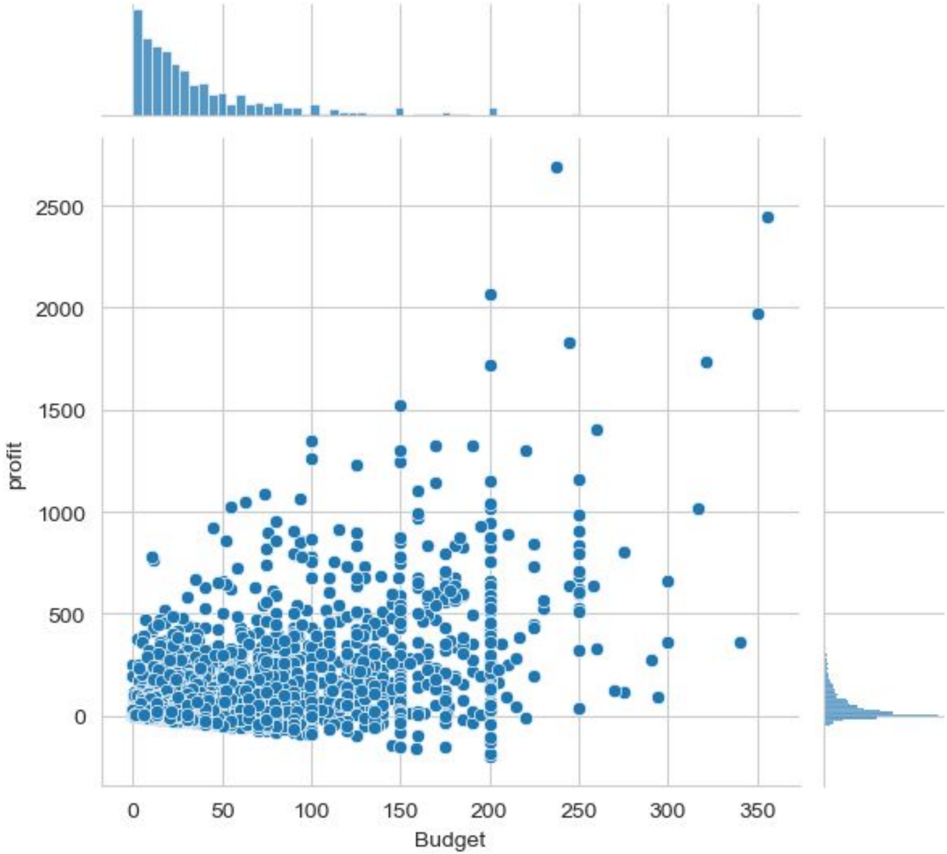
# Profit analysis

## Findings

### Observations

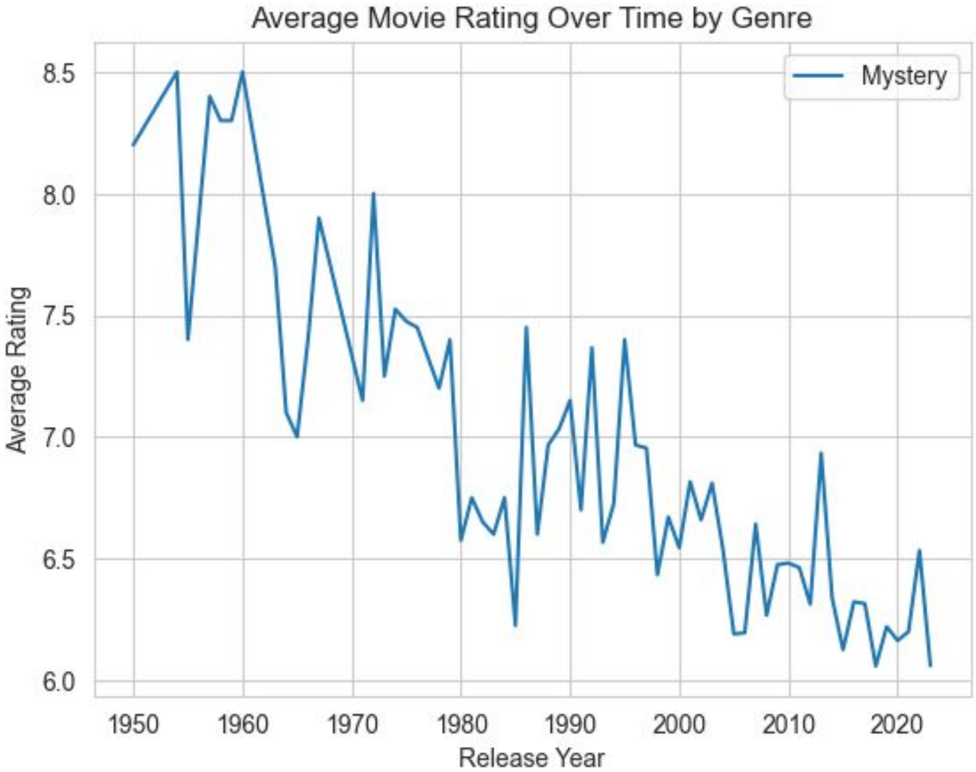
- Both Budget and profit are slightly skewed to right which means most of the movies had positive profit.

- There is a **positive relation** between profit and budget.



# Mystery's Decline: A Puzzle Losing Its Pieces

- The graph depicts a general decline in average ratings for the genre from the 1950s to 2020, with notable fluctuations and periods of resurgence.

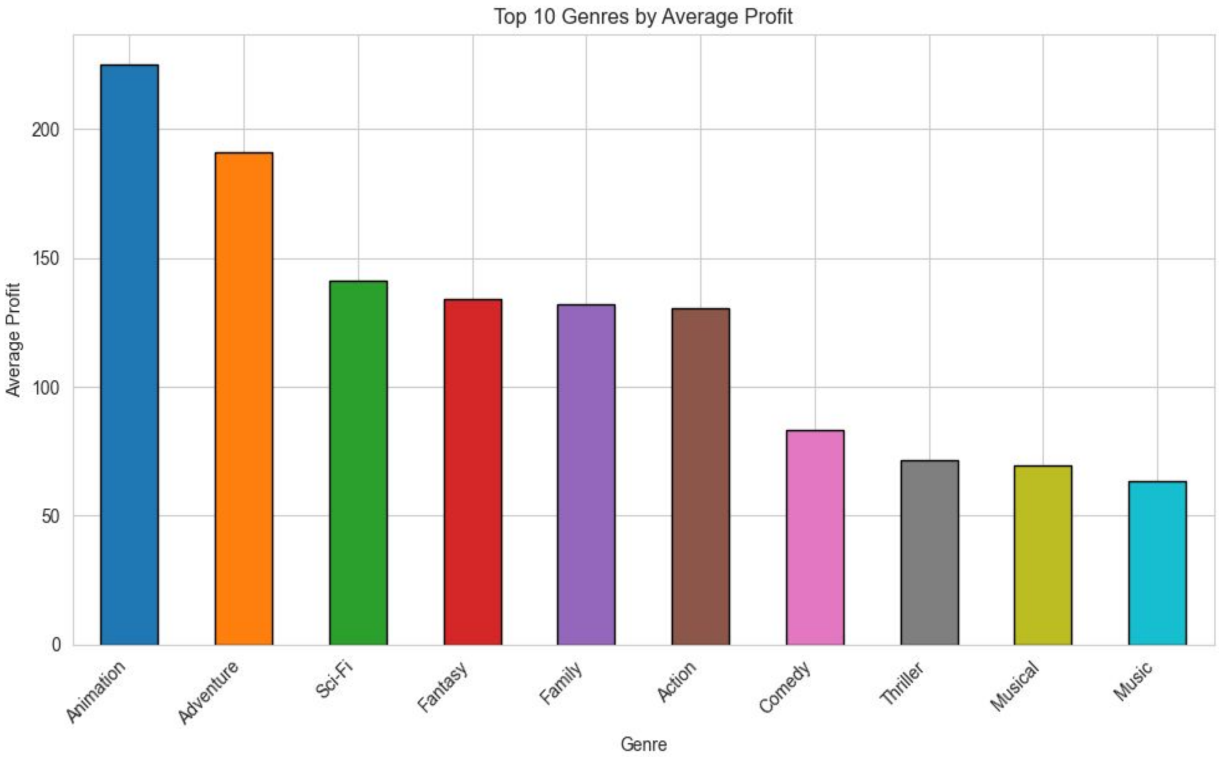


# Dad, Can We Go and See that Animation?

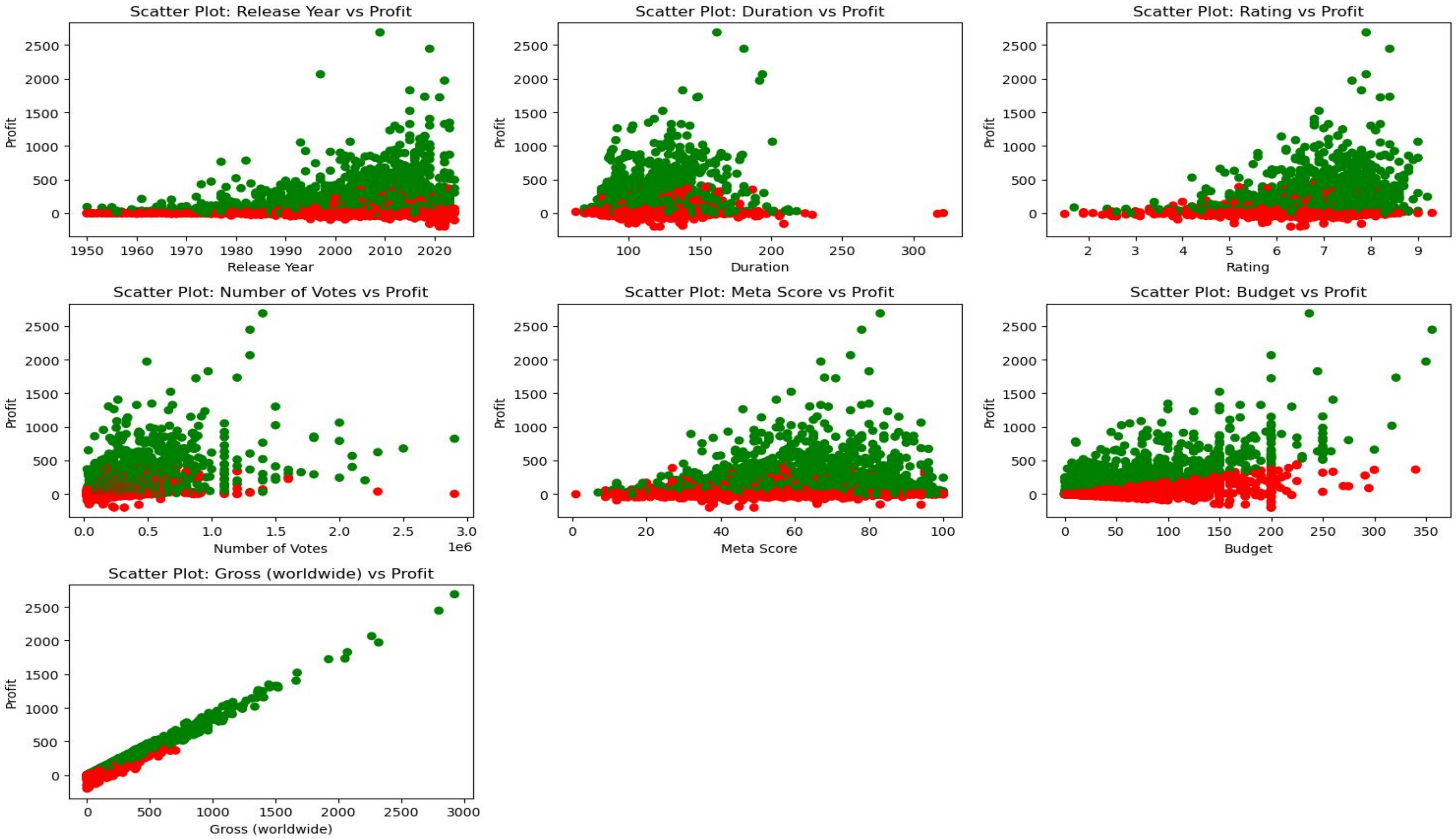
The bar chart titled ‘Top 10 Genres by Average Profit’ unveils the financial success stories of different movie genres.

Animation, leading the pack, outshines other genres, making it a popular choice for a cinema outing.

Adventure and Sci-Fi trail closely, while Music genre lags at the bottom



# Profitable and unprofitable movies properties



# Observations

1. Profit - release year

- The samples are skewed to the left which means as time passed directors are better in producing profitable movies.

2. Provit-Metascore

- We can see that at around 750 to 1000 million dollars, we get scores ranging from 40 to 90,

3. Profit-rating

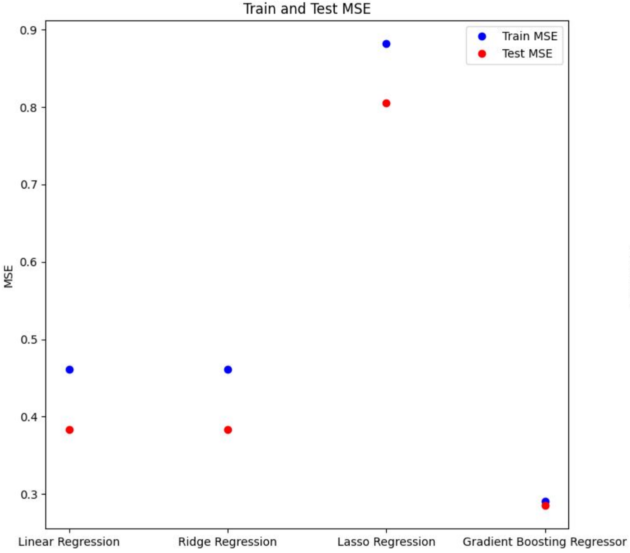
- Datums are skewed to the left which means If the rating is higher, there is a lower chance that the film is unprofitable

# Modeling The IMDB Using Neural Network

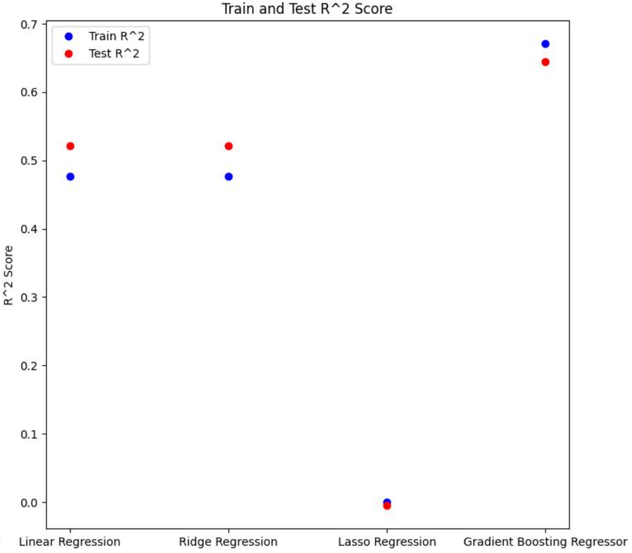
# Using Different types of Regression (No Neural Networks)

- Applied Linear, Ridge, Lasso, and Gradient Boosting regressions, plotting MSE and R².

- Gradient Boosting Regression outperforms neural network, achieving 0.67 R² on training and 0.64 on test data.



MSE

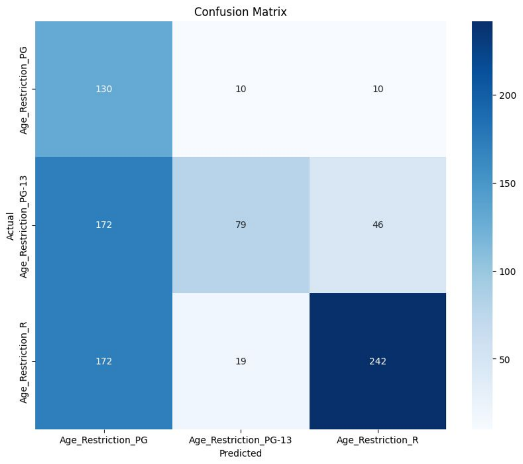


R² Score

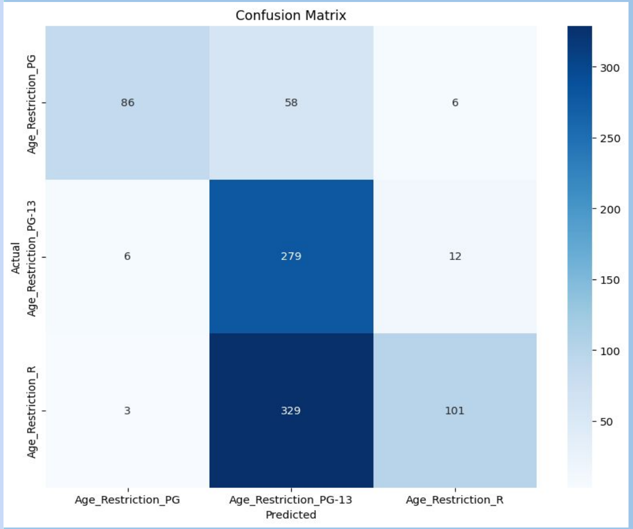
# Age Restriction Feature

| | |
|---|---|
| 1. PG | ● Parental Guidance Suggested |
| 2. PG-13 | ● Advised for age 13 and over |
| 3. R - Restricted | ● Children Under 17 Requires Accompanying Parent or Adult Guardian |

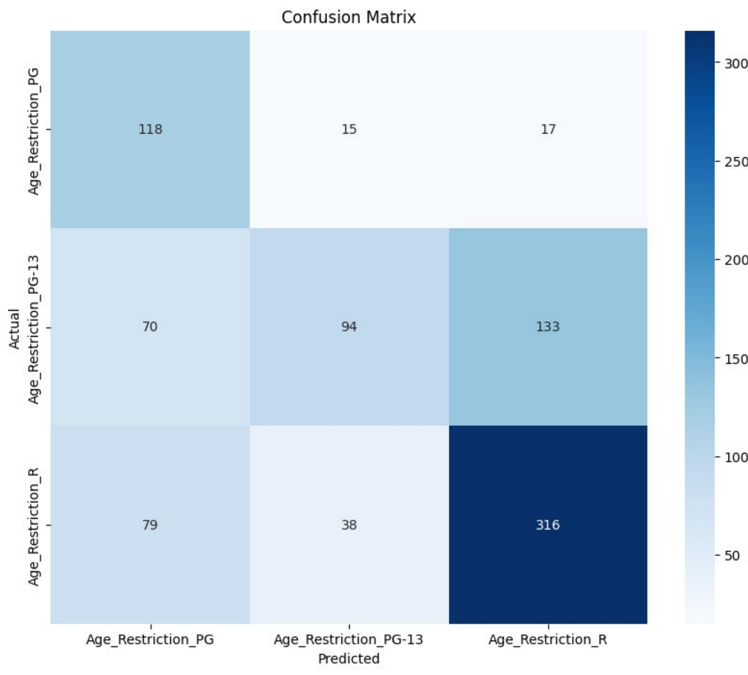
## *First Try*



## *Optimizing Parameters*



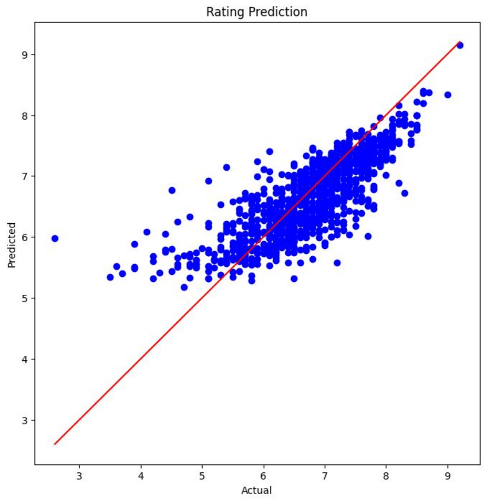
## Applying Smote for Balancing

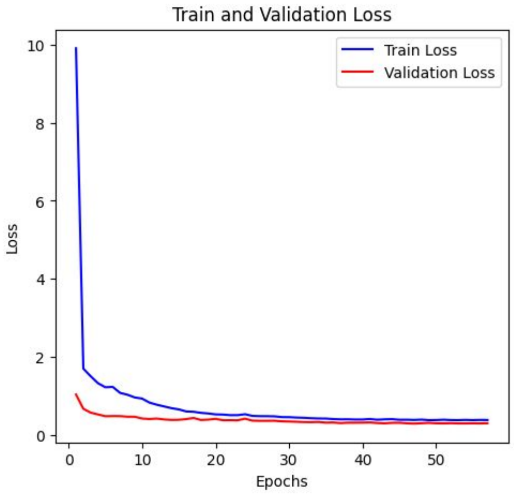


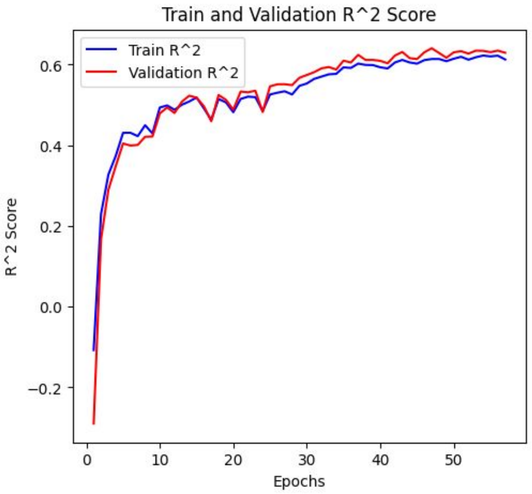
# Using Neural Networks for Regression Tasks

- The neural network got an $R^2$ score of around 0.62 on both the training and test data,

- So it predicts reasonably well but there's definitely room to improve.

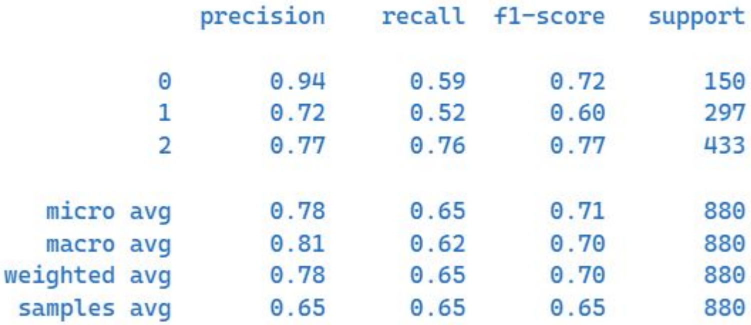

Rating Prediction


Train and Validation Loss

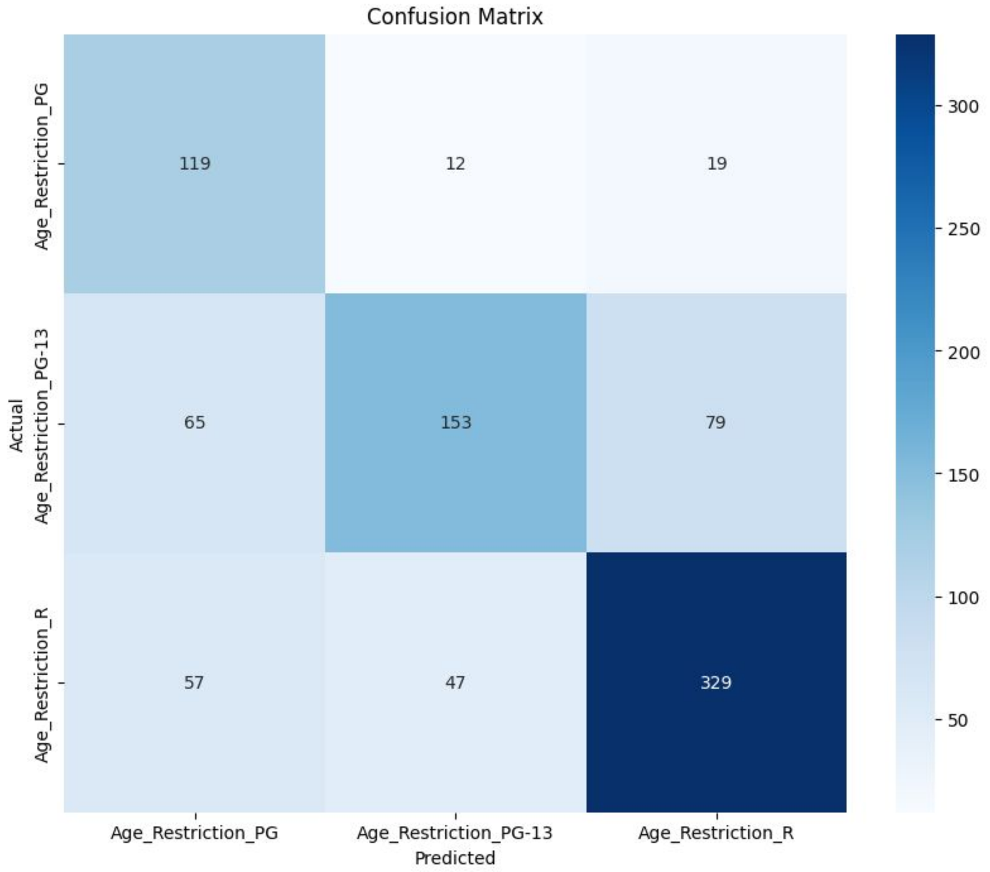

Train and Validation R^2 Score

# Modeling The IMDB Using Random Forest Classifier

# Random Forest Performance

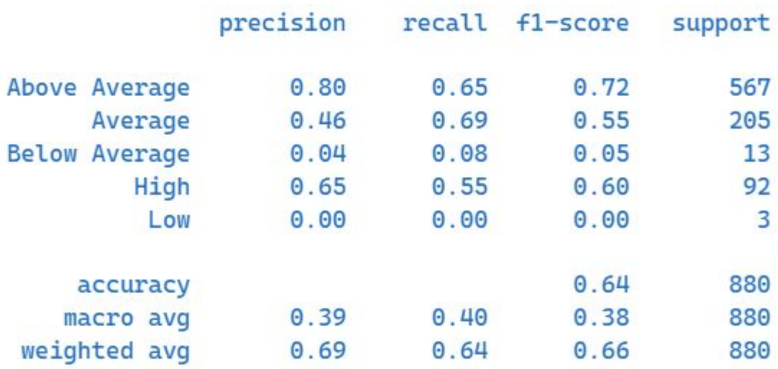


```
              precision    recall  f1-score   support

           0       0.94      0.59      0.72       150
           1       0.72      0.52      0.60       297
           2       0.77      0.76      0.77       433

   micro avg       0.78      0.65      0.71       880
   macro avg       0.81      0.62      0.70       880
weighted avg       0.78      0.65      0.70       880
 samples avg       0.65      0.65      0.65       880
```

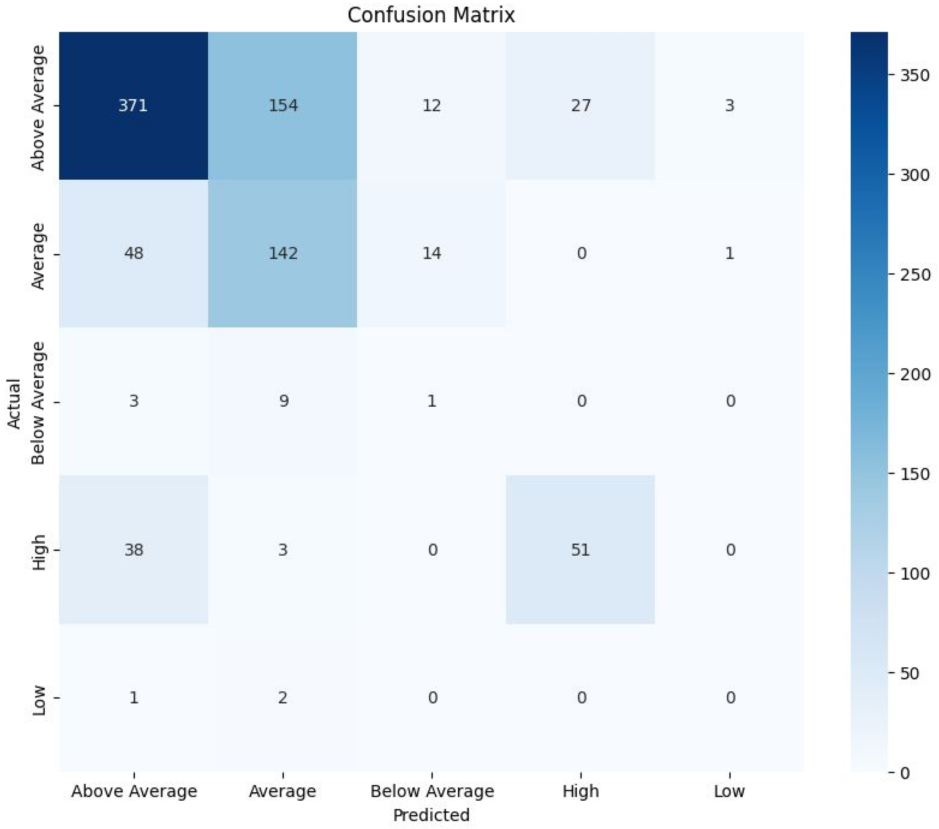


# Using Neural Network

# Rating Bucketing: Categorizing Ratings for Analysis

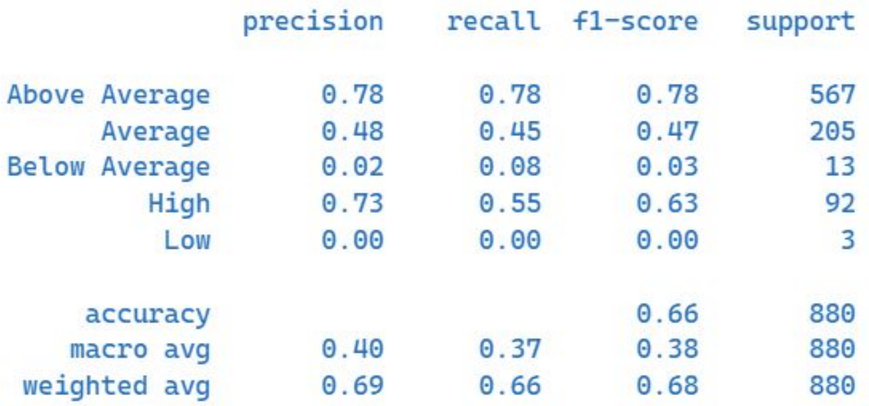


```
               precision    recall  f1-score   support

Above Average       0.80      0.65      0.72       567
      Average       0.46      0.69      0.55       205
Below Average       0.04      0.08      0.05        13
         High       0.65      0.55      0.60        92
          Low       0.00      0.00      0.00         3

     accuracy                           0.64       880
    macro avg       0.39      0.40      0.38       880
 weighted avg       0.69      0.64      0.66       880
```

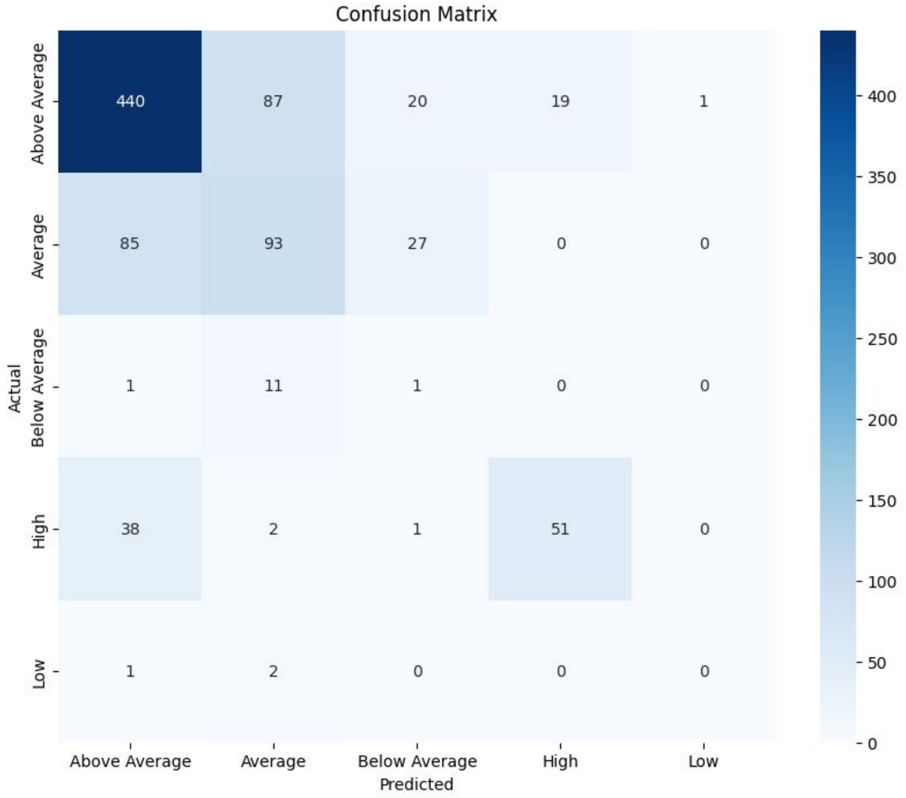


| Actual \ Predicted | Above Average | Average | Below Average | High | Low |
|---|---|---|---|---|---|
| Above Average | 371 | 154 | 12 | 27 | 3 |
| Average | 48 | 142 | 14 | 0 | 1 |
| Below Average | 3 | 9 | 1 | 0 | 0 |
| High | 38 | 3 | 0 | 51 | 0 |
| Low | 1 | 2 | 0 | 0 | 0 |

# Rating Bucketing: With balancing data



```
               precision    recall  f1-score   support

Above Average       0.78      0.78      0.78       567
      Average       0.48      0.45      0.47       205
Below Average       0.02      0.08      0.03        13
         High       0.73      0.55      0.63        92
          Low       0.00      0.00      0.00         3

     accuracy                           0.66       880
    macro avg       0.40      0.37      0.38       880
 weighted avg       0.69      0.66      0.68       880
```



Confusion Matrix

| Actual \ Predicted | Above Average | Average | Below Average | High | Low |
|---|---|---|---|---|---|
| Above Average | 440 | 87 | 20 | 19 | 1 |
| Average | 85 | 93 | 27 | 0 | 0 |
| Below Average | 1 | 11 | 1 | 0 | 0 |
| High | 38 | 2 | 1 | 51 | 0 |
| Low | 1 | 2 | 0 | 0 | 0 |

THANK YOUR ATTTENTION