# IMDb Top 250 Movies Analysis

By Taewon Jung, Seunggyun Shin, Cee Yun

# Introduction

**Data: IMDb Top 250 Movies Dataset**

**Source: Kaggle.com**

**Size: 250 x 20**

**Columns: rank, movie_id, title, year, link, imbd_votes, imbd_rating, certificate, duration, genre, cast_id, cast_name, director_id, director_name, writer_id, writer_name, storyline, user_id, user_name, review_id, review_title, review_content**

**Motivation: Extracting insights from different features of movies rated via IMDb users.**

Reference: https://www.kaggle.com/datasets/karkavelrajaj/imdb-top-250-movies/code

# Data Pre-Processing

- There were two NA values in the certificate column, and our group searched online and filled them out

```
imbd_votes      0
imbd_rating     0
certificate     2
duration        0
genre           0
```

```
data[(data["certificate"].isna()) | (data["certificate"] == "Unrated")]
#The Boat is Rated R
#Requiem for a Dream is Rated R
```

| | rank | title | year | imbd_votes | imbd_rating | certificate | duration | genre | self |
|---|---|---|---|---|---|---|---|---|---|
| **76** | 77 | The Boat | 1981 | 253,534 | 8.4 | NaN | 149 | Drama | 0 |
| **84** | 85 | Requiem for a Dream | 2000 | 852,453 | 8.3 | Unrated | 102 | Drama | 0 |

## Changed of Our Data Column

```
array(['rank', 'movie_id', 'title', 'year', 'link', 'imbd_votes',
       'imbd_rating', 'certificate', 'duration', 'genre', 'cast_id',
       'cast_name', 'director_id', 'director_name', 'writer_id',
       'writer_name', 'storyline', 'user_id', 'user_name', 'review_id',
       'review_title', 'review_content'], dtype=object)
```

```
array(['rank', 'title', 'year', 'imbd_votes', 'imbd_rating',
       'certificate', 'duration', 'genre', 'self', 'cast_n', 'review_n',
       'recent'], dtype=object)
```

# Descriptive Analytic Research Questions

1. **What is the relationship between the duration of movie and IMDb votes?**

- New column ‘recent’ - from ‘year’
- Summary statistic
  - The mean duration is longer for recent movies
  - Imdb votes is also higher for recent movies.

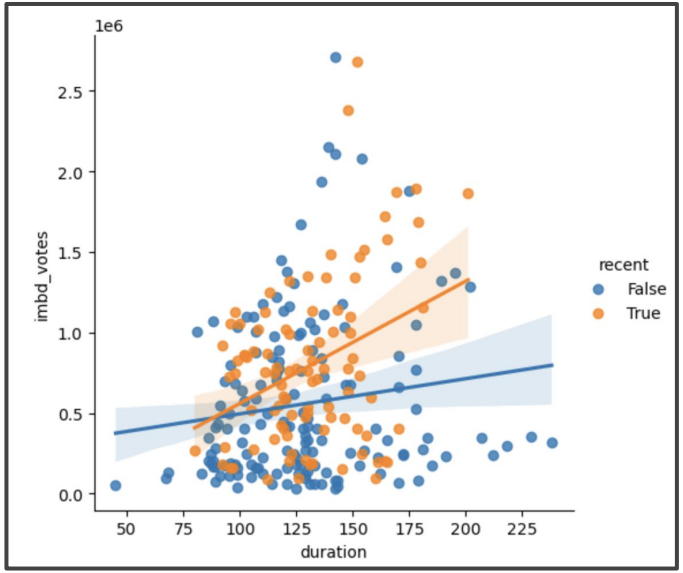
| recent | duration | imbd_votes |
|---|---|---|
| **False** | 127.941558 | 555848.12987 |
| **True** | 131.041667 | 795252.53125 |

# Descriptive Analytic Research Questions

**2. How do the duration and IMDb votes changes based on whether a movie was released recently(after 2000) or not?**

**FODS**

- **F**orm: linear
- **O**utlier: Some outliers (relatively long duration & very high votes)
- **D**irection: Both Positive (recent & older movies)
- **S**trengths: Moderately strong



# Conclusion

- **Scatter plot:**
  - Moderately strong, showing positive relationship between duration and imbd_votes for both recent and older movies.
  - There are some outliers with relatively long duration and very high imbd votes.
  - The center of the spread is around 100-135 with concentration around 0 and 1 million votes.
- **Summary statistic:**
  - The mean duration is longer for recent movies(after 2000) and imdb votes is also higher for recent movies.
- **Conclusion**
  - **Votes and duration of the recent movies(after 2000) are higher than the old ones.**

# Thank you!
# Any questions?