

# Give Me the Next **AAA** Title!

*An end-to-end Data Science project flow encompassing data sourcing, analysis, modeling, and storytelling using publicly-available IMDb.com data*

*Prepared By:*

**David Angelo Brillantes**

*dsbrillantes@up.edu.ph*

*Nov. 2 - 8, 2020*

# OUTLINE

## **I. Tickets, Please: Introduction**

- Overview of IMDB Data

## **II. Curtain Call: Data Sourcing**

- Data Modeling and Cleaning
- Web Scraping

## **III. Movie Magic: Exploratory Data Analysis**

- Identification and Problem Solving

## **IV. Cineaste: “Hit” Movie Metric Rationalization**

- Consultation & Interview

## **V. Lights, Camera, Action: Statistical Modeling**

- Building a Predictive Model

## **VI. That’s a Wrap: Conclusions & Recommendations**

# INTRODUCTION

- *The Internet Movie Database (IMDB)*
- *Overview of IMDB Data*

# THE INTERNET MOVIE DATABASE



...the world's most popular and authoritative source for movie, TV and celebrity content, designed to help fans explore the world of movies and shows.

*"We help you jog your memory about a movie, show, or person on the tip of your tongue, find the best movie or show to watch next, and empower you to share your entertainment knowledge and opinions with the world's largest community of fans."*

Statistics as of October 2020:

- 7,419,772 Titles
- 11,103,201 Personalities

# OVERVIEW OF IMDB DATA

Only a subset of the IMDB data is available for access and non-commercial use.

- Information on Titles
- Director and Writer Information
- TV Episode Information
- Principal Cast/Crew Information
- IMDB Ratings and Votes for Titles

# OVERVIEW OF IMDB DATA

Each dataset is contained in a TSV formatted file.

*Table 1. Datasets and Fields Summary*

DATASET	INFORMATION	FIELDS	USED
title.akas	“Also Known As” attributes	titleId, ordering, title, region, language, types, attributes, isOriginalTitle	Y
title.basics	basic info. for titles	tconst, titleType, primaryTitle, originalTitle, isAdult, startYear, endYear, runtimeMinutes, genres	Y
title.crew	director and writer information for titles	tconst, directors, writers	N
title.episode	TV episode information	tconst, parentTconst, seasonNumber, episodeNumber	N
title.principals	principal cast/crew for titles	tconst, ordering, nconst, category, job, characters	Y
title.ratings	IMDB rating and votes information for titles	tconst, averageRating, numVotes	Y
name.basics	basic info. for names	nconst, primaryName, birthYear, deathYear, primaryProfession, knownForTitles	Y



# OVERVIEW OF IMDB DATA

## A few notes about IMDB ratings:

- Individual votes are aggregated and summarized as a single IMDB rating.
- The rating displayed on each page is a **weighted average**, calculated using an undisclosed method.
- The ratings are “accurate” only in the sense that the formula used is consistent and unbiased. They provide a “fun and useful indication of the opinion of a movie held by the general public,” and act as more of a guide.
- External factors such awards and critic reviews are **not** considered.
- Safeguards to detect and stop rating inflation/deflation are used in the calculation.
- Ratings on the **Top 250 Rated** lists use a different formula, and only consider “regular voters,” who are identified with an undisclosed criteria.

# DATA SOURCING

- *Data Modeling & Cleaning*
- *Web Scraping*





# DATA SOURCING & CLEANING

- Data was retrieved on Nov. 3, 2020 from [datasets.imdbws.com](https://datasets.imdbws.com)
- UTF-8 Encoding where NULLs are encoded as '\N'
- Filtered to titles with type 'movie' and 'tvMovie'
- Filtered to titles with release year until 2020 only
- Dropped unnecessary columns
- Created other calculated columns
- Filtered other tables accordingly

# WEB SCRAPED DATA

## 1. ISO-3166-1 Codes for Countries

- Parsed through .JSON file
- **Goal:** Attach country names to title.akas
- Packages: `jsonlite`

# WEB SCRAPED DATA

## 2. Top 250 Rated Movies

- Scraped the IMDB web page
- **Goal:** Attach the corresponding tconst from title.basics
- Defined functions for getting:
  - Storyline & Plot Keywords
  - Release Date
  - Budget & Worldwide Gross
- Looped through each link; string manipulation; regex matching
- Packages: xml2, rvest, stringr, stringi

# WEB SCRAPED DATA

## 3. List of Movie Trilogies

- Scraped a Wikipedia article that contained a list of film series with three entries
- **Goal:** Transform to tidy list of series and entries including title and release year
- Found 673 series, 2019 entries
- Packages: `xml2`, `rvest`, `stringr`, `stringi`

# DATA MODELING

- All the tables were inserted into an **SQLite** database file.

Table 2. Working Datasets

DATASET	INFORMATION	FIELDS
imdb_top250	Top 250 Rated Movies	Rank, Title, ReleaseYear, IMDB_Rating, tconst, ReleaseDate, Plotline, Keywords, Budget, WW_Gross
title_basics	basic info. for titles	tconst, titleType, primaryTitle, isAdult, startYear, runtimeMinutes, mainGenre*, subGenre1*, subGenre2*, tconst_int*, NumGenres*, FilmAgeYears*
title_aka	"Also Known As" attributes	titleId, title, language, isOriginalTitle, country_name*
title_prin	principal cast/crew for titles	tconst, ordering, nconst, category
title_rating	IMDB rating and votes information for titles	tconst, averageRating, numVotes
name_basics	basic info. for names	nconst, primaryName, birthYear, deathYear, knownForTitles, CastAgeYears*, mainProfession*, subProfession1*, subProfession*
wiki_trilogies	movie trilogies	SeriesId, Series, Entry_1, Entry_2, Entry_3

\*calculated field

# EXPLORATORY DATA ANALYSIS

■ *Identification & Problem Solving*





# EDA

## 1. What is the highest rated movie in 2018?

- Tables Used: title\_basics, title\_rating
- Some movies got a perfect score of 10, but with  $<10$  votes
- Create a new variable, **Popularity**, that serves as an index calculated by multiplying the average rating with the number of votes
- Split the dataset into two: movies that are below and above the median Popularity
- Look at the top-rated movies within those two groups
- Filter the dataset to those with number of votes above the median number of votes

# EDA

- Highest Rated are Documentaries and Dramas
- Still relatively unknown with a small number of votes

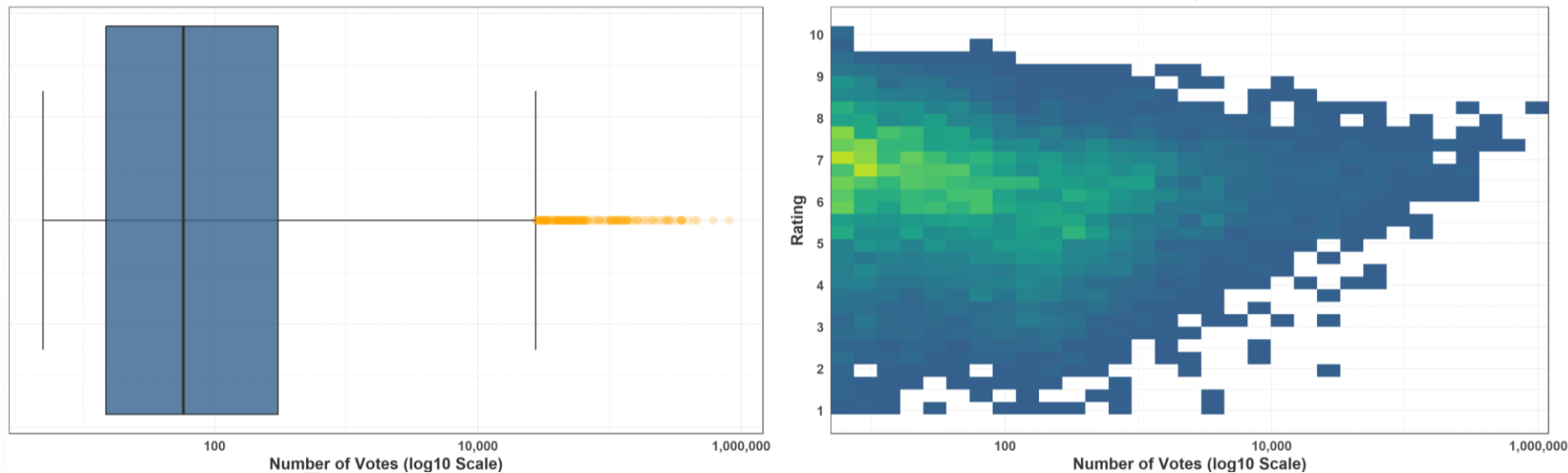
*Table 3. Three Highest Rated Movies of 2018 by Popularity*

TITLE	GENRE	RATING	VOTES	POPULARITY
HIGH POPULARITY				
The Nagano Tapes: Rewound, Replayed & Reviewed	Documentary	9.3	221	2,055.3
The Transcendents	Music	9.2	1,456	13,395.2
Eghantham	Drama	9.2	644	5,924.8
LOW POPULARITY				
Life: Amidral	Drama	9.7	70	679.0
Filthy Frank Final Full Lore Movie	Comedy	9.6	112	1,075.2
Manry at Sea: In the Wake of a Dream	Documentary	9.5	75	712.5

# EDA

- Ratings are heavily skewed to the right by very popular movies.
- Most of the ratings are in the 6 to 9 range.

Figure 1. Boxplot of # of Votes (L) and 2D Distribution of # of Votes and Ratings (R)



# EDA

- Another approach to this question is to use the **same definition** that IMDB uses in curating their **Top Rated Movies** list, which shows the top rated 250 movies
- the “formula provides a true ‘Bayesian estimate’, which takes into account the number of votes each title has received, minimum votes required to be on the list, and the mean vote for all titles.”

$$Top\ 250\ Rating = \left( \frac{v}{v + m} \cdot R \right) + \left( \frac{m}{v + m} \cdot C \right)$$

Where,

- $v$  is the number of votes for the movie
- $m$  is minimum votes required to be listed in the Top Rated list (currently 25,000)
- $R$  is the average rating of the movie, and
- $C$  is the mean vote across all titles

# EDA

- The only pre-filtering we do is to consider only movies with at least 25,000 votes.
- The highest rated movie of 2018 is **Avengers: Infinity War**.

Table 4. Three Highest Rated Movies of 2018 by “Top Rating”

TITLE	GENRE	RATING	VOTES	TOP RATING
HIGH POPULARITY				
Avengers: Infinity War	Action	8.4	809,261	8.345605
Spider-Man: Into the Spider-Verse	Action	8.4	347,298	8.278111
Green Book	Biography	8.2	351,834	8.092846
LOW POPULARITY				
K.G.F: Chapter 1	Action	8.2	32,961	7.503338
They Shall Not Grow Old	Documentary	8.3	27,838	7.488478
Tumbbad	Drama	8.3	25,570	7.452082

# EDA

## 2. Who was the most popular actor in 2018?

- Tables Used: title\_prin, name\_basics
- We look at both actors and actresses who assumed the **leading roles** in their films.
- Ways to measure actor/actress popularity:
  - Number of pageviews to profile
  - Number of listed credits
  - Frequency of appearance in user search queries
  - **Popularity of projects**
- Make use of movie data and use ratings and votes as proxies for popularity.
- Create a new variable, **Spotlight**, calculated as the sum of Popularity divided by the number of movies he/she was in, summarized for each Actor/Actress.



# EDA

- The most popular actor and actress in 2018 were **Robert Downey Jr.** and **Emily Blunt**. IMDB's 2018 Top 10 Stars list (based on pageviews) is noticeably different.

*Table 5. Five Most Popular Lead Actors/Actresses in Movies released in 2018*

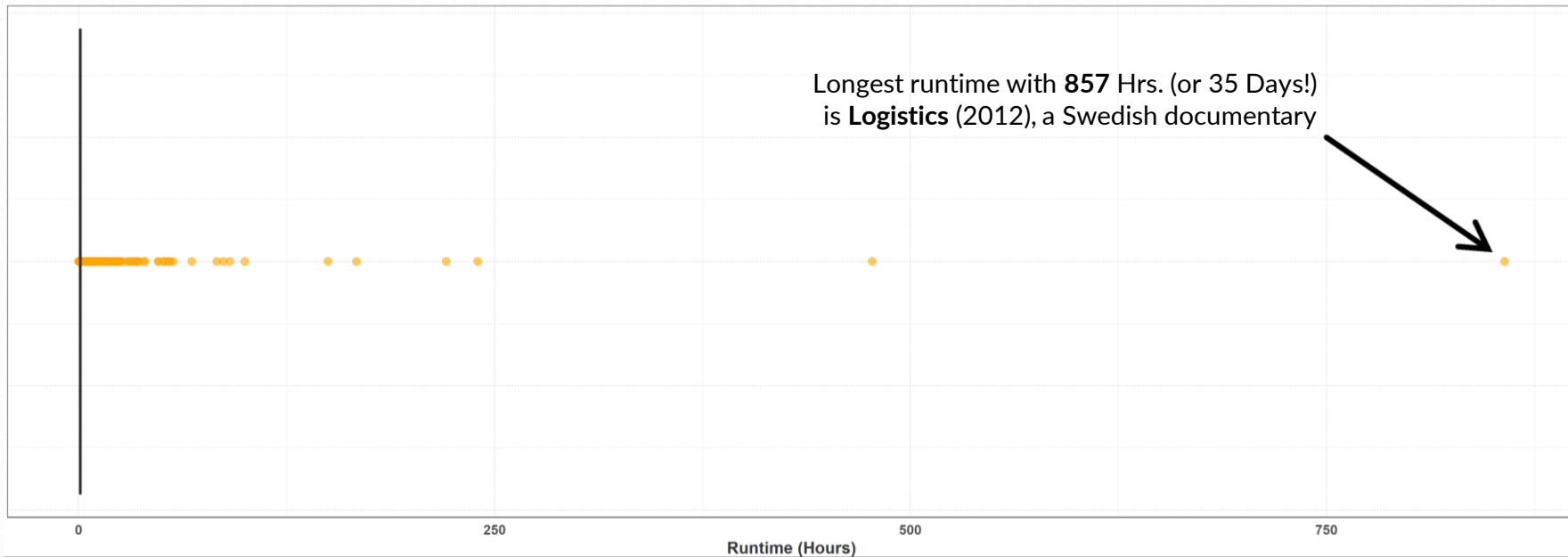
NAME	TITLES	SPOTLIGHT
ACTOR		
Robert Downey Jr.	Avengers: Infinity War	6,797,792
Chadwick Boseman	Black Panther	4,470,396
Ryan Reynolds	Deadpool 2	3,593,675
Rami Malek	Bohemian Rhapsody	3,511,472
Shameik Moore	Spider-Man: Into the Spider-Verse	2,917,303
ACTRESS		
Emily Blunt	Mary Poppins Returns, A Quiet Place	1,742,585
Toni Collette	Hereditary	1,676,730
Olivia Colman	The Favourite	1,250,108
Cate Blanchett	Ocean's 8	1,226,598
Jennifer Lawrence	Red Sparrow	1,063,682

# EDA

## 3. What is the longest runtime of a single movie?

■ Tables Used: title\_basics

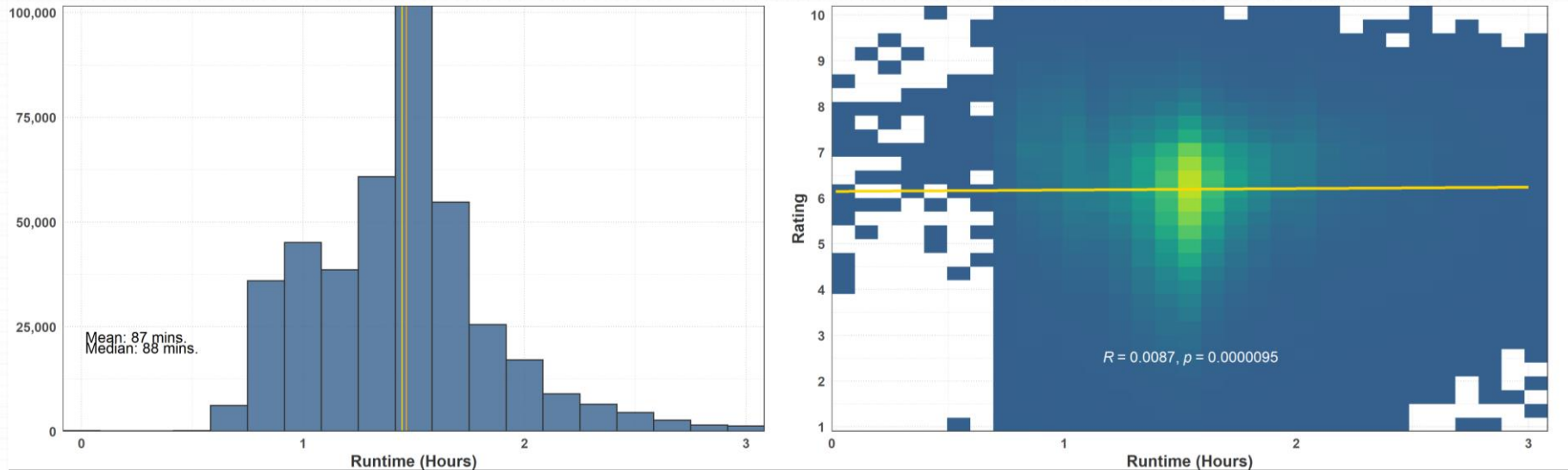
Figure 2. Boxplot of Movie Runtimes



# EDA

- Filtering to movies  $\leq 3$  Hrs. long, most have a runtime between 1 to 2 Hrs.
- There is no correlation between the runtime and the rating of the movie.

Figure 3. Histogram of Movie Runtime (L) and 2D Distribution of Movie Runtime and Ratings (R)



# EDA

## 4. What is the longest aggregate runtime of a trilogy?

- Tables Used: title\_basics, wiki\_trilogies
- Identifying movies as being part of a trilogy was a challenge.
- No clear way to match; used titles and release years for exact string matching.
- Issues with exact matching:
  - Small differences in title formats
  - Small inconsistencies in release year
  - Titles within a series are often very similar to each other
  - IMDB primaryTitle could be different from Wikipedia
  - Considered only IMDB titles of type 'movie' and 'tvMovie'
  - Filtered the runtime table to those with non-null Runtime





# EDA

## 4. What is the longest aggregate runtime of a trilogy?

- A more involved approach: measure **String Similarity** using algorithms
  - Used both the **Levenshtein** (LV) distance and the full **Damerau-Levenshtein** (DL) distance methods as they are appropriate for our situation.
  - Set thresholds to consider as a valid match (e.g. .80, .95)
  - Considered the distance between the years where we allot a maximum discrepancy of 2 years.
  - **Normalized** the strings to an extent (removed whitespaces, changed to uppercase, changed Roman numerals to Arabic)
  - Also considered 'video' type titles aside from 'movie' and 'tvMovie'
  - Looked at the AKA information for fails

# EDA

## 4. What is the longest aggregate runtime of a trilogy?

- Defined functions for normalizing strings, measuring similarity, and a wrapper
- Normalized the target variables, primaryTitle, originalTitle, and all akaTitle
- Applied matching function in a strictest  laxest fashion, varying parameters:
  - String Sim. Threshold (  is stricter, from 0.70 to 0.95)
  - Year Diff. Threshold (  is stricter, from 0 to 2)
  - Top  $n$  Similar Strings (  is stricter, from 1 to 2)
- Used 7 sets of parameters, going through each until a match is found
- Searched through 1,784,309 target titles



# EDA

- Matching on title and year only
- **1183 were successfully matched**

Table 6. First Attempt at Matching (Exact String Match)

SERIES ID	SERIES	ENTRY #	TITLE	YEAR	TCONST
1	9½ Weeks	Entry_1	9½ Weeks	1986	tt0091635
1	9½ Weeks	Entry_2	Love in Paris	1997	NA
1	9½ Weeks	Entry_3	The First 9½ Weeks	1998	NA
2	12 Rounds	Entry_1	12 Rounds	2009	tt1160368
2	12 Rounds	Entry_2	12 Rounds 2: Reloaded	2013	tt2317524
2	12 Rounds	Entry_3	12 Rounds 3: Lockdown	2015	tt3957956
3	2012	Entry_1	2012: Doomsday	2008	NA
3	2012	Entry_2	2012: Supernova	2009	NA
3	2012	Entry_3	2012: Ice Age	2011	NA

# EDA

- Spot checked through fails to confirm if not in database
- **1909 final matches**

Table 6. Second Attempt at Matching (String Similarity Search Algorithm)

SERIES	TITLE	NORMALIZED TITLE	YEAR	FINAL MATCHING RESULT
9½ Weeks	9½ Weeks	9½WEEKS	1986	tt0091635
9½ Weeks	Love in Paris	LOVEINPARIS	1997	originalTitle_target Love in Paris tt0119576
9½ Weeks	The First 9½ Weeks	THEFIRST9½WEEKS	1998	primaryTitle_target The First 9 1/2 Weeks tt0164026
12 Rounds	12 Rounds	12ROUNDS	2009	tt1160368
12 Rounds	12 Rounds 2: Reloaded	12ROUNDS2:RELOADED	2013	tt2317524
12 Rounds	12 Rounds 3: Lockdown	12ROUNDS3:LOCKDOWN	2015	tt3957956
2012	2012: Doomsday	2012:DOOMSDAY	2008	originalTitle_target 2012: Doomsday tt1132130
2012	2012: Supernova	2012:SUPERNOVA	2009	primaryTitle_target 2012: Supernova tt1479847
2012	2012: Ice Age	2012:ICEAGE	2011	primaryTitle_target 2012: Ice Age tt1846444

# EDA

- The movie trilogy with the longest aggregate runtime is the **Flash Gordon** serials followed by the **Trilogy of Henryk Sienkiewicz**.

Table 7. Top Five Trilogies with Longest Aggregate Runtime

SERIES	TITLES	AGG. RUNTIME (MINS.)	AGG. RUNTIME (HRS.)
Flash Gordon (serials)	Flash Gordon (1936), Flash Gordon's Trip to Mars (1938), Flash Gordon Conquers the Universe (1940)	764	12.73
Trilogy of Henryk Sienkiewicz	With Fire and Sword (1999), The Deluge (1974), Colonel Wolodyjowski (1969)	622	10.37
The Human Condition	No Greater Love (1959), Road to Eternity (1959), A Soldier's Prayer (1961)	579	9.65
Once Upon a Time Trilogy	Once Upon a Time in the West (1968), Duck, You Sucker! (1971), Once Upon a Time in America (1984)	551	9.18
The Godfather	The Godfather (1972), The Godfather Part II (1974), The Godfather Part III (1990)	539	8.98

# EDA

## 5. Have there been changes in trend in genres in the past 10 years?

- Tables Used: title\_basics, title\_ratings
- Filtered to movies released starting in 2011
- Looked at **trend** within each genre in terms of:
  - Number of movies
  - The average rating
- 27 unique genres consisting of: Comedy, Drama, Documentary, Horror, Biography, Adventure, Animation, Action, Sci-Fi, Thriller, Crime, Mystery, Romance, Fantasy, Musical, Family, Music, Adult, History, Sport, War, Western, Reality-TV, Talk-Show, News, Game-Show, and Short

- For (L): # of Movies are topped by Documentaries and Dramas, and the rankings are stable
- For (R): Ranking have changed a lot; newly popular compared to 2011 are History, Biography, and Comedy

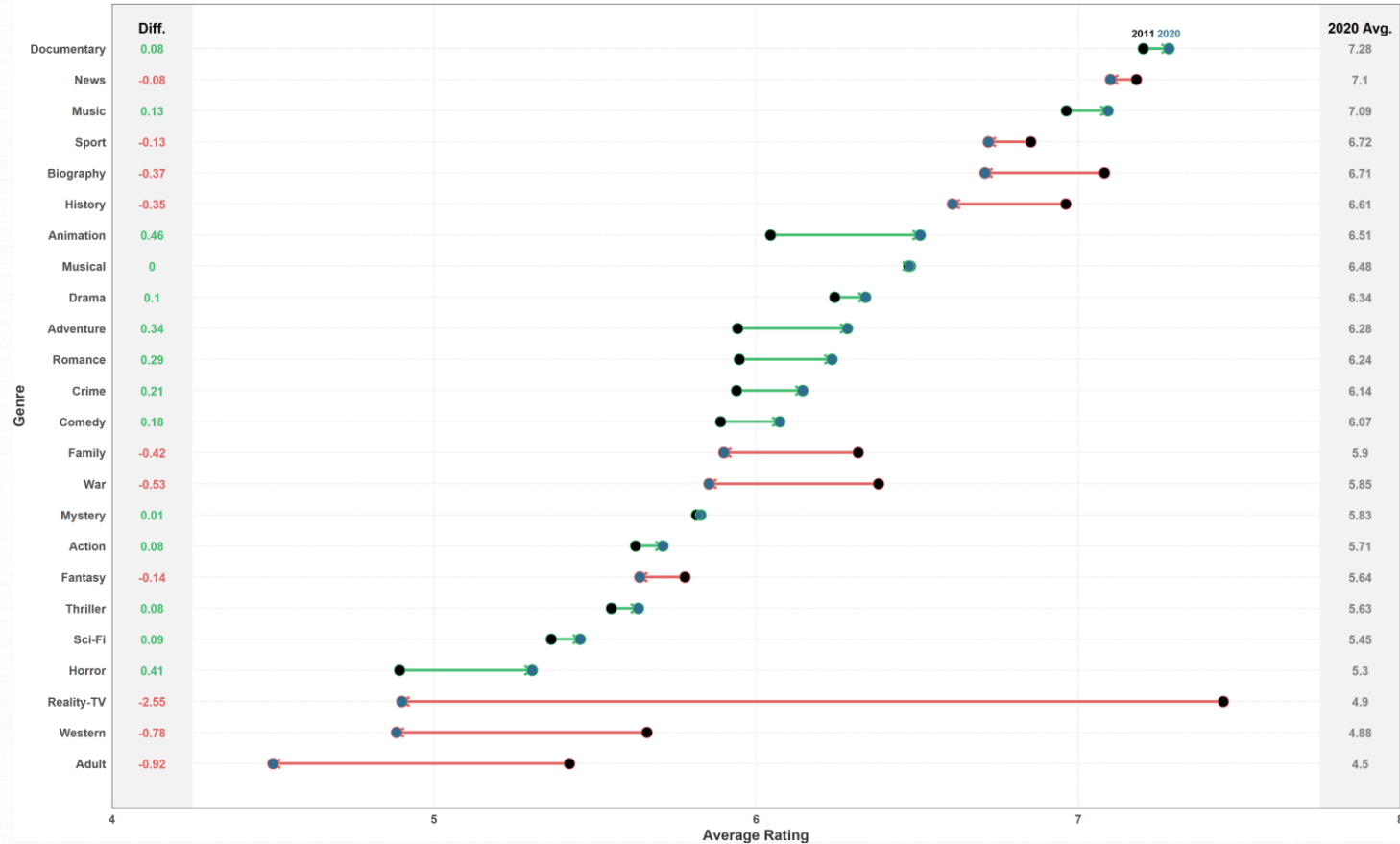
Figure 4. Rankings of Genre by # of Movies (L) and Rankings by Popularity weighted by # of Movies





The Genre with the highest rating in 2020 is **Documentary**. Reality-TV had largest decrease while Animation had the largest increase.

Figure 5. Average Rating by Genre in 2011 and in 2020

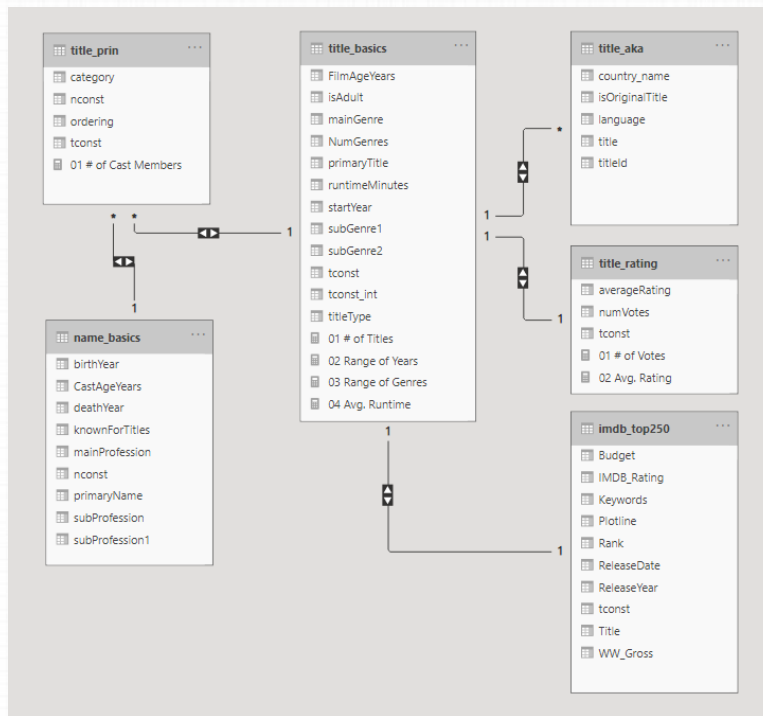




# EDA

## 6. Create a dashboard to visualize the dataset

- Used Power BI



# **“HIT” MOVIE RATIONALIZATION**

■ *Consultation and Interview*

# DEFINING A BUSINESS METRIC

- Consulted and interviewed those knowledgeable about the industry



**Josam Dalman**

*UPD Film Major  
EVP, UP Cineaste's Studio  
Filmmaker*



**Jog Dela Peña**

*UPD Visual Communications Major  
UP Photography Society  
Digital Artist*

# DEFINING A BUSINESS METRIC

- Questions to gather insights were constructed.
- Q#2: *In your opinion, what percentage of a film's "success" can be attributed to each of viewership, ratings, and earnings (e.g. 20%-40%-40%, 30%-50%-20%), and why?*
- Chose the **first component** from Principal Components Analysis incorporating multivariate data on votes, ratings, and.
- Weights from the answers in Q#2 will be used to make **two** definitions of the response,  $Y_1$  and  $Y_2$ .
- A **third** definition of equal weights was used to make  $Y_3$ .
- Finally, the values of each was averaged into the final composite index, *MetaScore*.
- Aimed to take the different perspectives into consideration.

# STATISTICAL MODELING

■ *Building a Predictive Model*



# STATISTICAL MODELING

**GETTING FINAL  
DATASET  
AND FEATURES**

**BUILDING MODELS**

**PREDICTING “HIT”  
MOVIES**



# MODEL DATASET

- Features came from the other IMDB tables.

Table 8. Feature Descriptions

FEATURE	CATEGORY	DESCRIPTION	SAMPLE VALUE
lead_age	Numeric	Age of lead actor/actress when movie was released	54
runtime_mins		Runtime of the movie in minutes	181
num_countries		Number of countries movie was released	43
num_languages		Number of languages movie was released	8
num_genres		Number of genres	3
main_genre	Categorical	The main genre of the movie	Action
lead_category		The category of the lead cast	Actor
years_from_2020		Number of years between release and 2020	1

# MODEL DATASET

- Other Dataset pre-processing:

1. Considered only movies released starting in 2016
2. Removed rows with missing values
3. Filtered to movies with at least 10,000 votes
4. Web scraped the Budget and Worldwide Gross of each movie
5. Budget values in foreign currency were converted to US dollars using the exchange rates given by the quantmod package

- The final dataset had **638** movies

# MODEL DATASET

## ■ In Creating *MetaScore*:

1. Values of VOTES, RATING, and EARNING were scaled to be between 0 to 100

2. The weights were then used to create  $Y_1$ ,  $Y_2$ , and  $Y_3$  via the formula,

$$Y_i = (\omega_{i1} \cdot VOTES) + (\omega_{i2} \cdot RATING) + (\omega_{i3} \cdot EARNING)$$

where  $\omega$  are the weights for the  $i^{\text{th}}$  definition of  $Y_i$

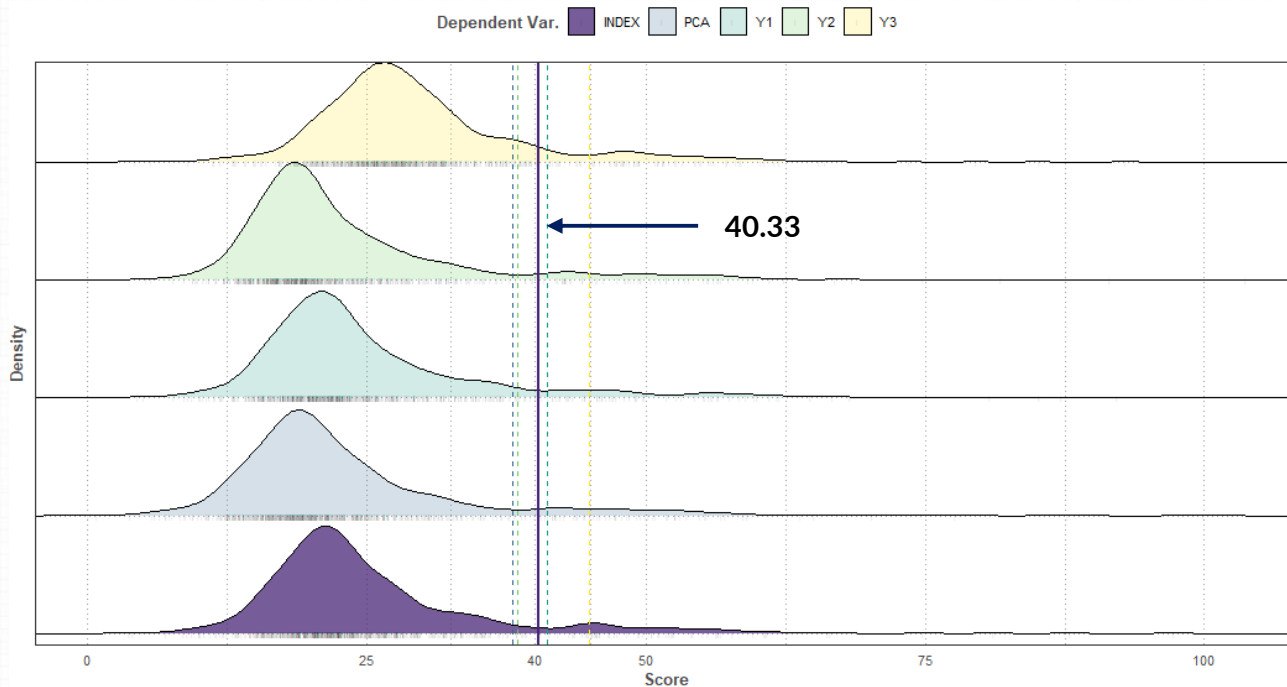
3. The observation coordinates taken from the first component of the PCA which explained 65% of the variance between the three variables were also scaled from 0 to 100

4. The index *MetaScore* is then the average of the four values

# MODEL DATASET

- For getting the final class labels for classification, a movie is considered a “Hit” when the *MetaScore* is greater than or equal the upper limit of the Tukey Outlier Fence ( $Q3 + 1.5 \cdot IQR$ ).

Figure 6. Empirical Densities of Dependent Variables (Vertical Lines = Tukey Outlier Fences)



# MODEL BUILDING

- The dataset was split into 80% training and 20% test sets

Table 9. Final Logistic Model after Stepwise Selection and Removing Non-Significant Variables

FEATURE	COEFFICIENTS	EXPONENTIATED COEFFICIENTS	P-VALUE
runtime_mins	0.0460	1.0471	0.0000
num_countries	0.2123	1.2635	0.0000
num_languages	0.1958	1.2163	0.0243
lead_category	-0.91120	0.4020	0.0300

Table 10. Goodness of Fit of Candidate Models

MODEL	RESIDUAL DEVIANCE	LRT P-VALUE
Full Model	188.18	-
Final Model	205.34	0.3095

# MODEL BUILDING

- A threshold probability of 0.5 was initially used

Table 11. Cross Tabulation of Actual and Predicted Labels

PREDICTED VALUES	ACTUAL VALUES	
	NOT HIT	HIT
NOT HIT	112	4
HIT	6	6

- The overall prediction accuracy is 0.9219 (95% CI: 0.861, 0.9619)
- Sensitivity: 0.6000
- Specificity: 0.9491
- **Positive Predictive Value: 0.5000**
- Negative Predictive Value: 0.9655



# MODEL BUILDING

- A higher threshold of 0.8 was checked to improve PPV

Table 12. Cross Tabulation of Actual and Predicted Labels

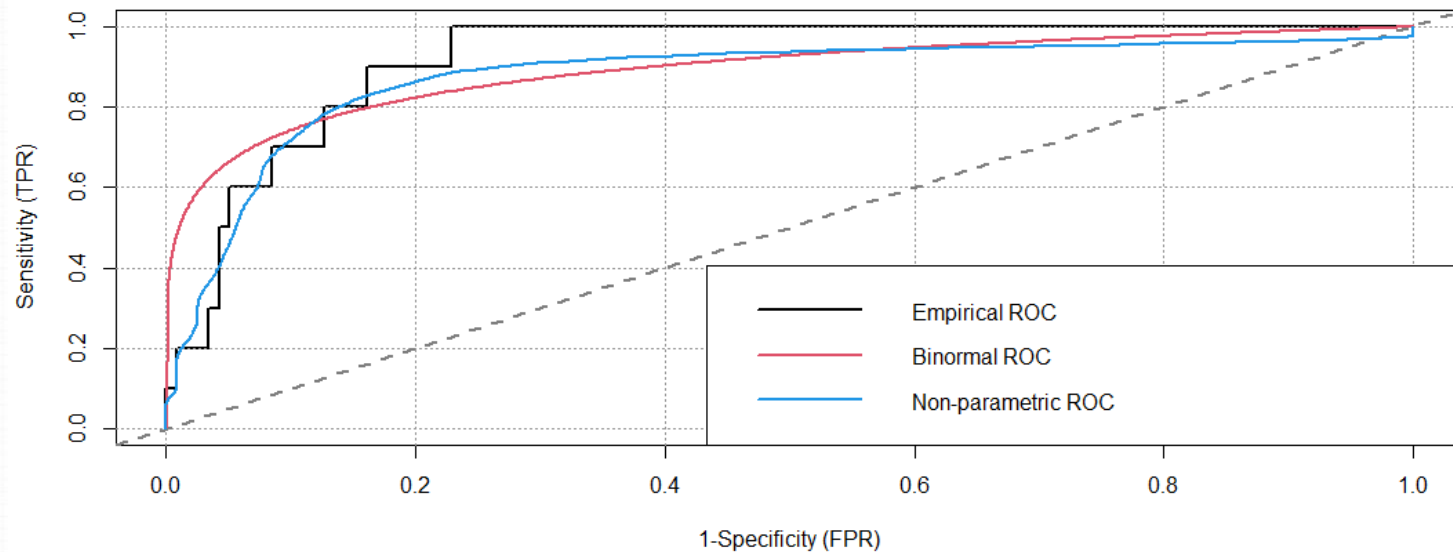
PREDICTED VALUES	ACTUAL VALUES	
	NOT HIT	HIT
NOT HIT	117 (+5)	8 (+4)
HIT	1 (-5)	2 (-4)

- The overall prediction accuracy is 0.9297 (95% CI: 0.8707, 0.9673)
- Sensitivity: 0.2000
- Specificity: 0.9915
- **Positive Predictive Value: 0.6667**
- Negative Predictive Value: 0.9360

# MODEL BUILDING

- The Area under the ROC Curve is 0.922

Figure 6. ROC Curve



# **CONCLUSIONS & RECOMMENDATIONS**

# CONCLUSIONS

- The IMDB dataset is considerably robust for analysis.
- We created a composite index that scores a movie's success using data on votes (viewership), rating, and monetary earnings.
- We were able to build a Logistic Regression model for predicting a "Hit" movie with a PPV of 0.67.
- Consultation with experts reveal the following:
  - The essential characteristics of a great movie are Technicality, Consistency, and Identity vs. Form and Content
  - Weights given were 50-25-25 and 30-20-50 (Votes-Ratings-Earning)
  - The future of the film industry is unpredictable, but they wish to see: more originals, and continue giving directors creative control on their works.

# RECOMMENDATIONS

- There is still a lot to explore with the existing IMDB data.
- The data can be enriched by looking at other existing datasets that are publicly available (e.g. OpenLens' MovieLens data)
- The model can still be improved by including more complex features that encapsulate the public's opinion on the movies, considering ratings by other established movie-review platforms, and characterizing financial success in other areas.
- The model can be combined with other algorithms and ML methods to create an ensemble model.
- It's worth exploring the possibility of creating a robust inference model to find driving variables towards the success of a movie.
- A review of the existing literature is needed to further concretize the results of the study.

# Give Me the Next **AAA** Title!

*An end-to-end Data Science project flow encompassing data sourcing, analysis, modeling, and storytelling using publicly-available IMDb.com data*

Thank you for listening!