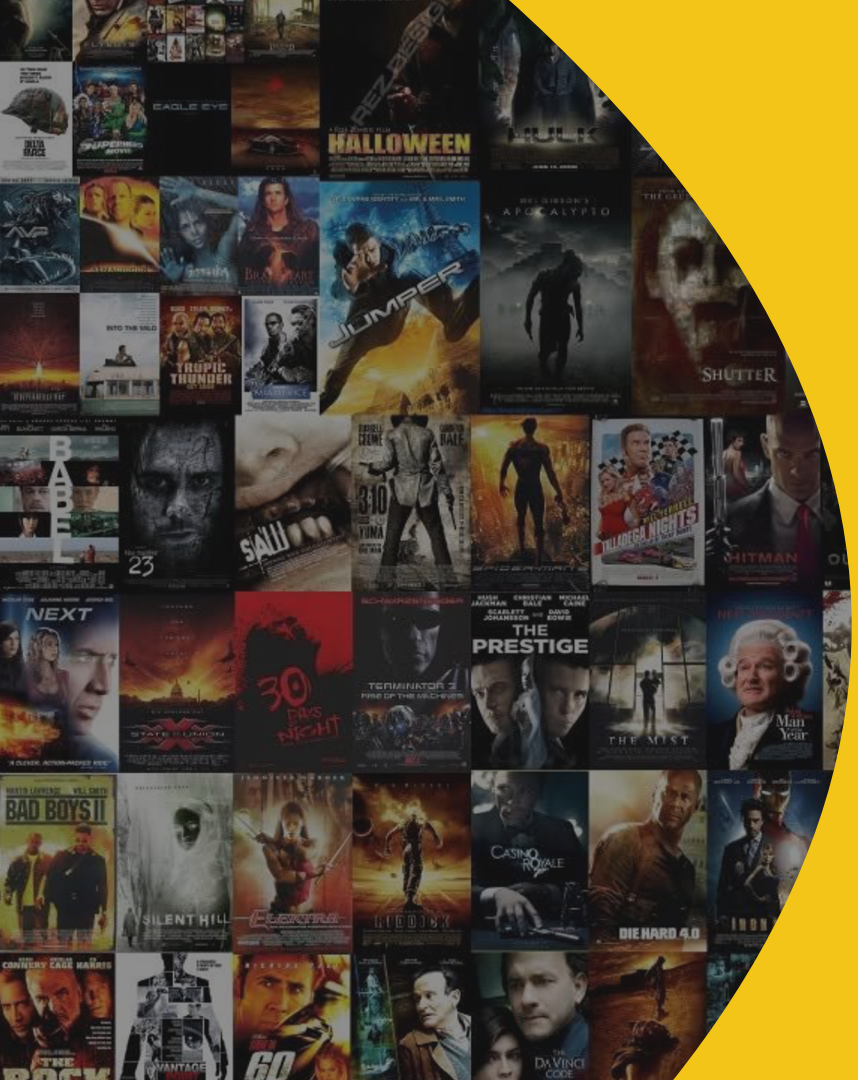


BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

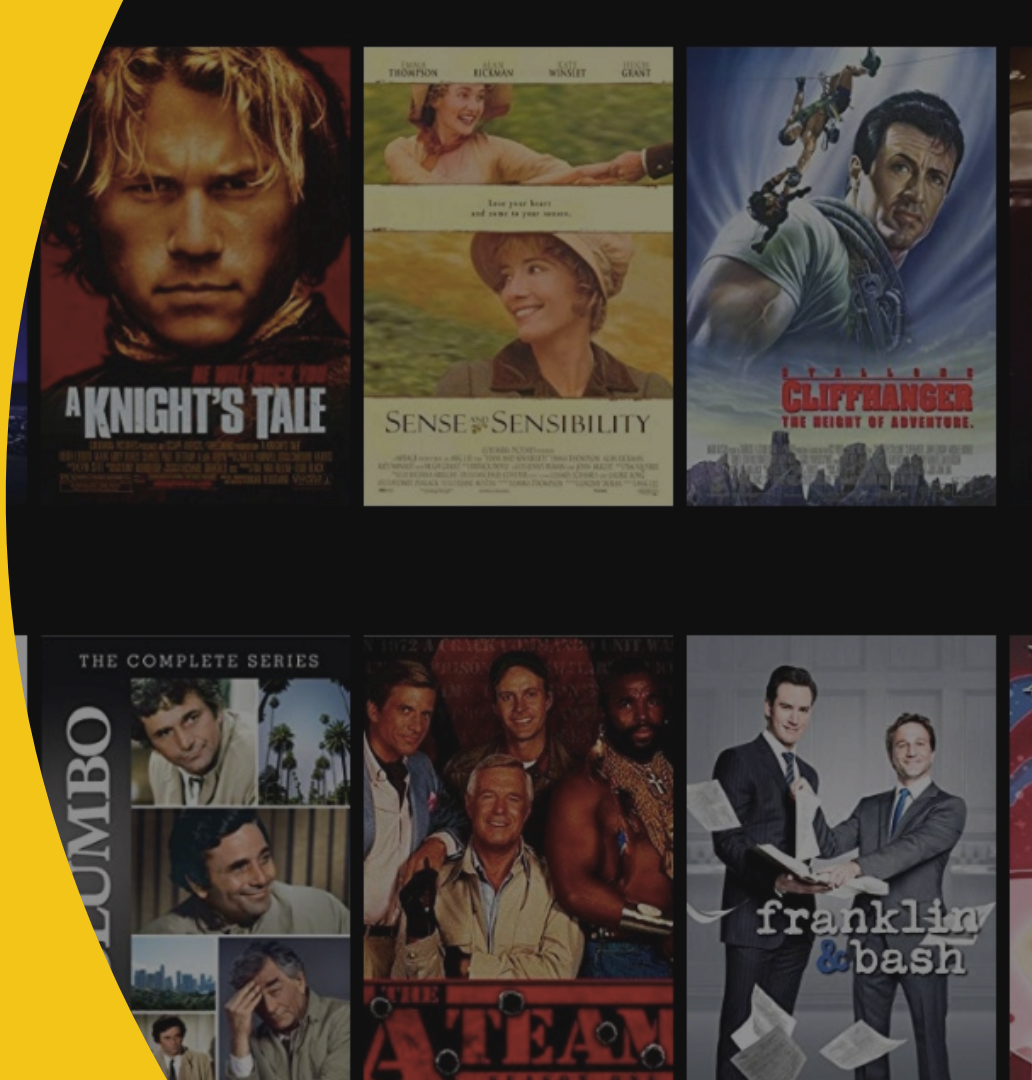
Digital Technologies and Value Creation

IMDb Movie Database Analysis

Han Tongova
Matheus Diaz-Maciel
Anchenmaa Kara-Ool
Sirada Chuntararuangnapa

Agenda

1. Firm and Question Selection
2. Data Collection Method
3. Challenges
4. Dataset Lexicon
5. Visualisations
6. Potential Additional Data



Firm and Question Selection

- We selected **Universal Pictures** as the firm interested in the data analysis.
- Hypothetical scenario:
  - Universal Pictures is in the planning stage of producing a new blockbuster movie. It thus requires in-depth data analysis of the most popular blockbusters of the recent decades in order to understand the industry and the business scenario.
- Questions:
  - **What are the key characteristics of a popular movie?**
  - **How can Universal Pictures achieve the next Blockbuster?**

Data Collection Method

- IMDb is the movies and TV shows reliable data hub.
- We filtered movie ranged from 2000 until 2021 as the data is most relevant to the current productions and future filmmaking decisions of Universal Pictures.

- The data to be collected:
  - Release year
  - Movie title
  - User rating
  - Metascore
  - Vote
  - Director
  - Genre
  - Gross Income
  - Motion Picture Rating (MPAA)

Data Collection Method (Cont)

- To carry out the data collection, the method selected was to utilise a combination of the Selenium/WebDriver, Requests and BeautifulSoup Python libraries to scrape data from the IMDb database, filtered by the years aforementioned.

- The choice of Selenium allowed for the code to run across multiple pages in the IMDb website, enabling the collection of a larger amount of movie data. The sole use of BeautifulSoup and Requests would not allow this due to the nature of the IMDb website and how it loads content - dividing it across pages.

Challenges

What we found:

- Some of the information required was not properly displayed in the database, with small changes in the page's HTML code altering how the information was displayed.
- The website was divided into multiple pages.
- Numbers return in data frame are different dtype.
- The code often failed to compile director names for movies with more than one director.

How we overcame:

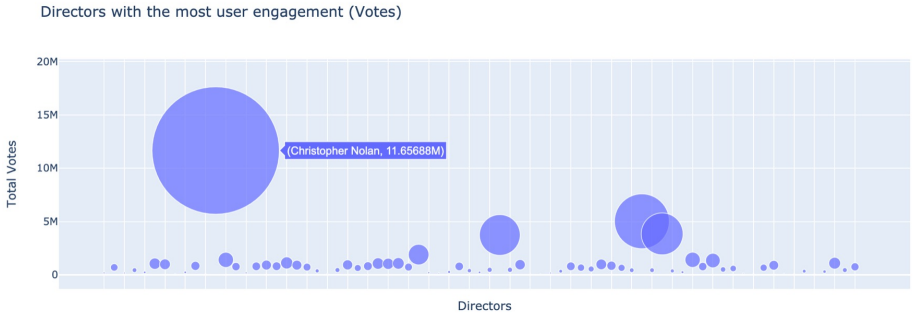
- Exceptions were implemented in the code to account for these potential changes.
- Creating a *while-loop to* devise a code that would iterate and compile information from all the pages.
- Changing all numbers to numerical (float64) before calculating the statistical summary
- Manual imputation the names of these directors.

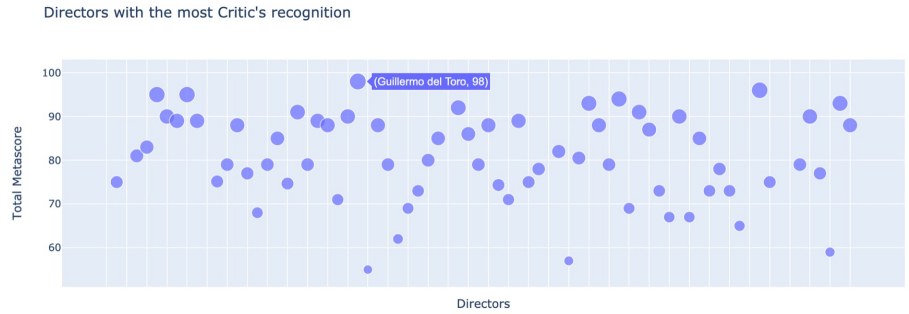
Dataset Lexicon

Variable	Description	Unit
Title	Movie title	Title
Year	Movie release year	Year
Critics Rating	Movie rating based on IMDb user's opinions and input	Score from 0-10
Vote	Vote count per movie	Number of vote
Director	Name(s) list of director(s) of each movie	Name(s)
Genre	Genre(s) of the movie	Genre(s)
Gross [millions]	Total revenue generated by the movie in US and Canada	Millions of Dollars
Motion picture rating (MPAA)	Age recommendation as defined by the MPAA (Motion Picture Association of America): G - General Audiences PG - Parental Guidance Suggested PG-13 - Parents Strongly Cautioned R - Restricted NC-17 -- No One 17 and Under Admitted	MPAA age recommendation scale
Metascore	Movie rating based on critic's opinions on the MetaCritics platform - A specialist platform for critics to evaluate movies and shows	Score from 0-100

	count	mean	std	min	25%	50%	75%	max
Critics rating	96.0	8.3	0.3	8.1	8.1	8.2	8.4	9.6
Metascore	88.0	80.0	9.9	55.0	73.8	80.0	88.0	98.0
Vote	96.0	689992.3	496151.1	65662.0	297073.5	660321.5	942883.5	2439902.0
Gross [millions]	85.0	129.6	163.5	0.0	11.3	59.1	188.0	858.4

Visualisations



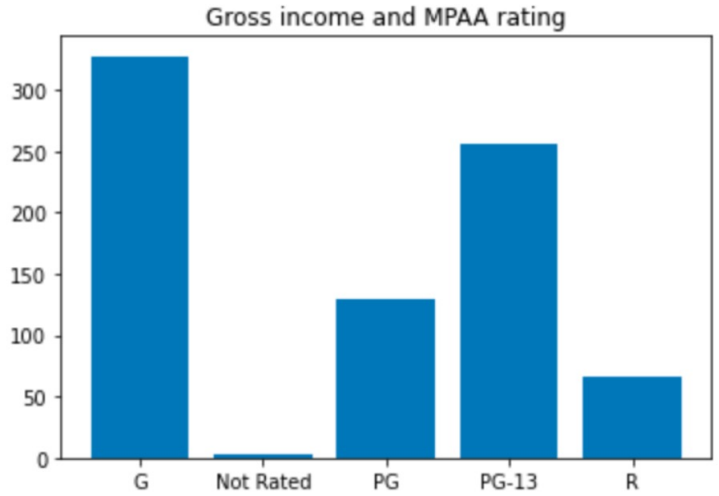


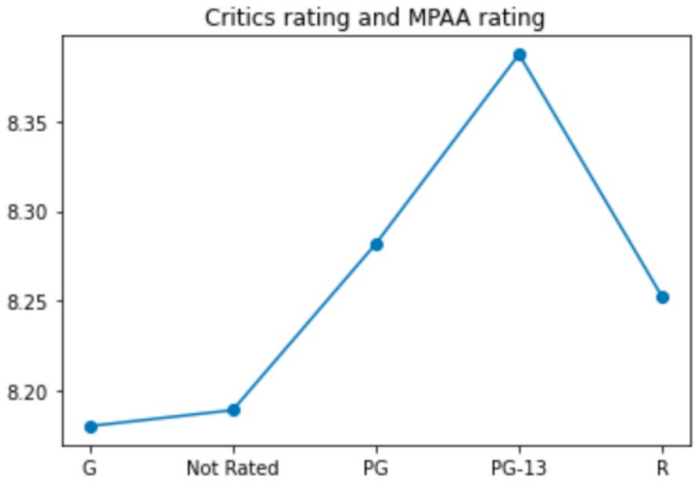
- Votes show how the general population felt about the films.
  - Directors with top Votes:
    1. Christopher Nolan 11.6 million
    2. Peter Jackson 5 million
    3. Quentin Tarantino 3.8 million
    4. Martin Scorsese 3.7 million

- Metascore is professional reviews from critics.
  - Directors with top Metascore:
    1. Guillermo del Toro 98 score
    2. Steve Mcqueen 96 score
    3. Andrew Stanton 95 score
    4. Asghar Farhadi 95 score

Visualisations

- G-rated or general films such as Finding Nemo, WALLE, and Monsters, Inc. earned the highest gross income in the United States and Canada at around 327 million, more adult rated films earned less.
- The reasons: the movies can be enjoyed by a wide range of audiences, and even though the main target is children, they are unable to attend the theatre alone.
- Rating-wise, all MPAA movies are nearly identical.



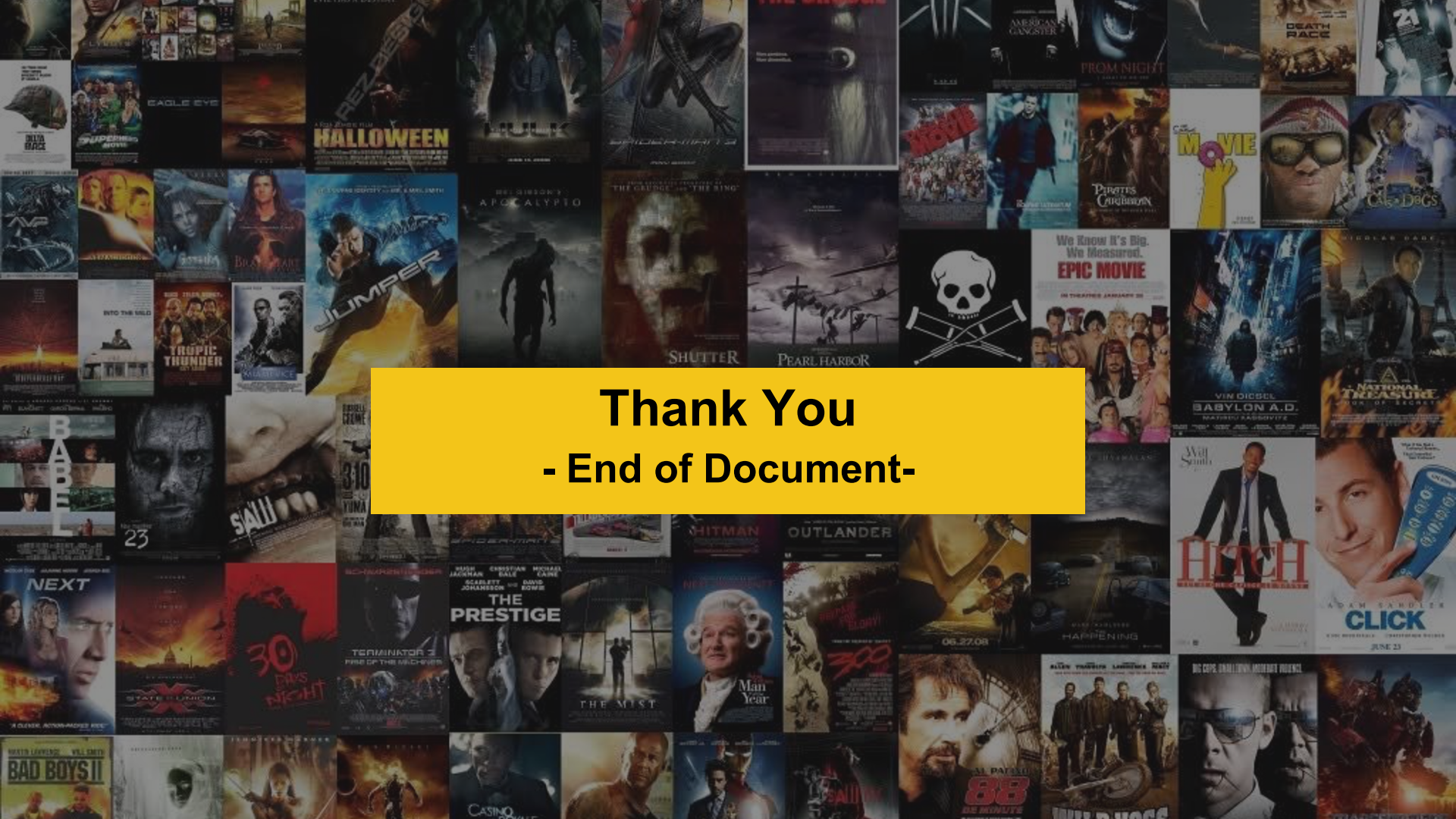


Potential Additional Data

- Data from IMDb reviews can provide some ideas and guidelines for what kind of movies Universal Pictures should make next, as well as which director and MPAA can be decided.
- However, the total amount of data was insufficient.
- Additional data required to complete the questions are:
  - Production budget
  - Marketing budget

  → The quality of production and promotion
  - Release month
  - Release week of month

  → The best time to release a film
  - Star names → Celebrity influence
  - Genres → Movie trends
  - Relevant # on social media → Online response of the movie



Thank You

- End of Document-