Team Oscar

December 6, 2018

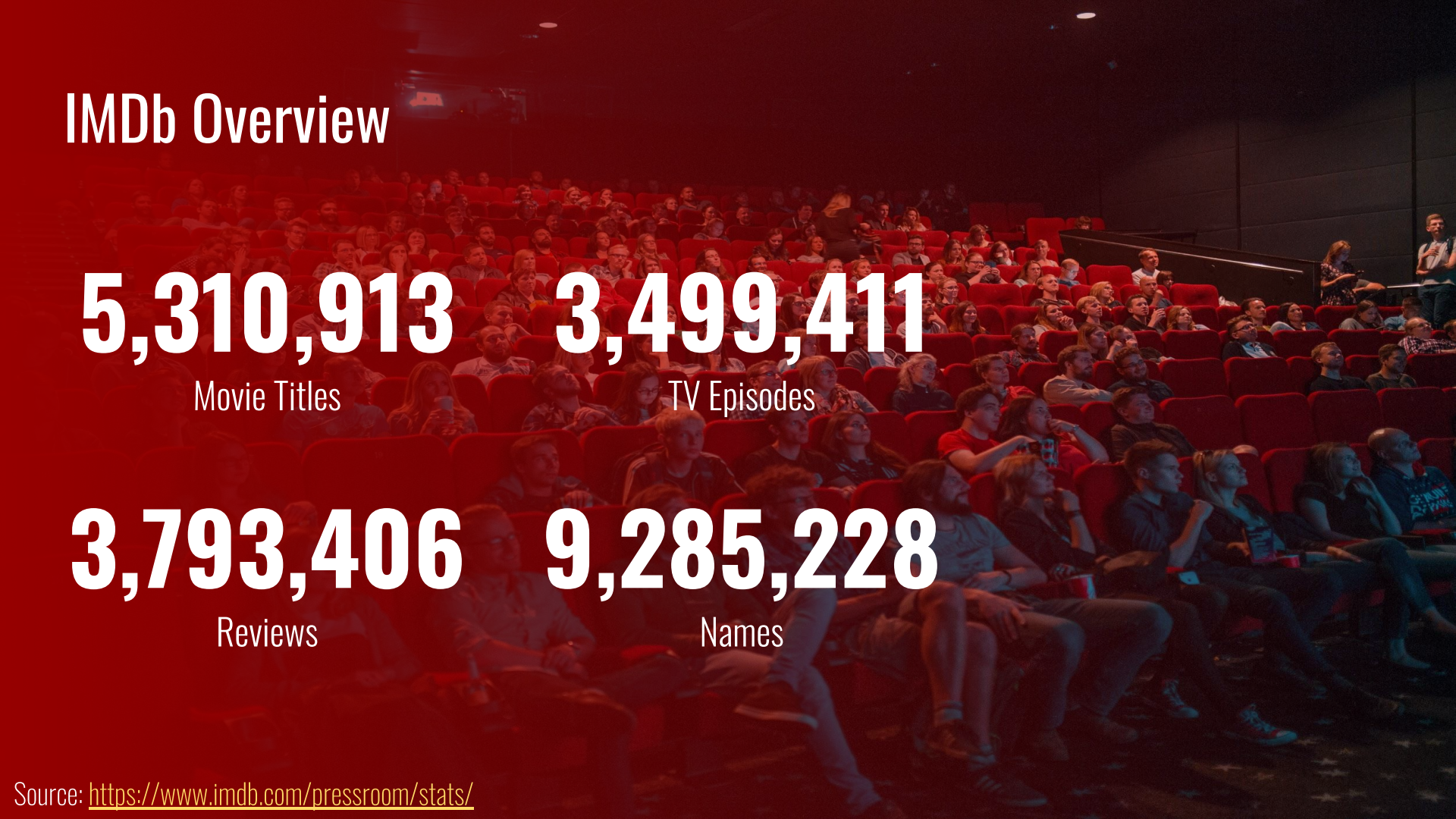
•••

Ann Eitrheim
Justin Ishikawa
Jamie Olds

Executive Summary

- IMDb contains over **5.3** million titles and over **9.3** million personalities
  - We are working with a subset of this data to build project Oscar
- Use case: Content sites need data to make smart purchasing decisions
  - Our solution pulls together this data for easy retrieval
- Project: Team Oscar proposes developing a SQL Database that allows easy access to search and analyze descriptive data about films
- We are using a combination of tools including R, MySQL, and Tableau

IMDb Overview



5,310,913
Movie Titles

3,499,411
TV Episodes

3,793,406
Reviews

9,285,228
Names

Source: https://www.imdb.com/pressroom/stats/

Design Considerations

Data Preparation

- Data had to be parsed into separate fields where arrays were saved as strings.
- We eliminated data that did not pertain to the overall goal of our project.
- Numbers saved as strings were converted into numeric data types.

Tools

- R and Python was used to parse and clean data for export to CSV.
- CSVs were exported using R to be uploaded into Google SQL where our data will be hosted due to size and accessibility.

tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000617	short	The Robber's Sweetheart	Røverens brud	0	1907	\N	\N	Drama,Short
tt0000618	short	Salaviinanpolttajat	Salaviinanpolttajat	0	1907	\N	20	Comedy,Short
tt5446528	movie	Mad Dog - From Chaos to Comeback	Mad Dog	0	2016	\N	72	Documentary,Drama,Sport
tt1659337	movie	The Perks of Being a Wallflower	The Perks of Being a Wallflower	0	2012	\N	103	Drama,Romance
tt5446602	tvMiniSeries	Extremos da Cidade	Extremos da Cidade	0	2014	\N	\N	Reality-TV

Business Use Case

Netflix made $11.6 billion in revenue last year through its streaming content services. In order for this project to capture some of that market share, we'll need to leverage our database to predict smart content investments based on...

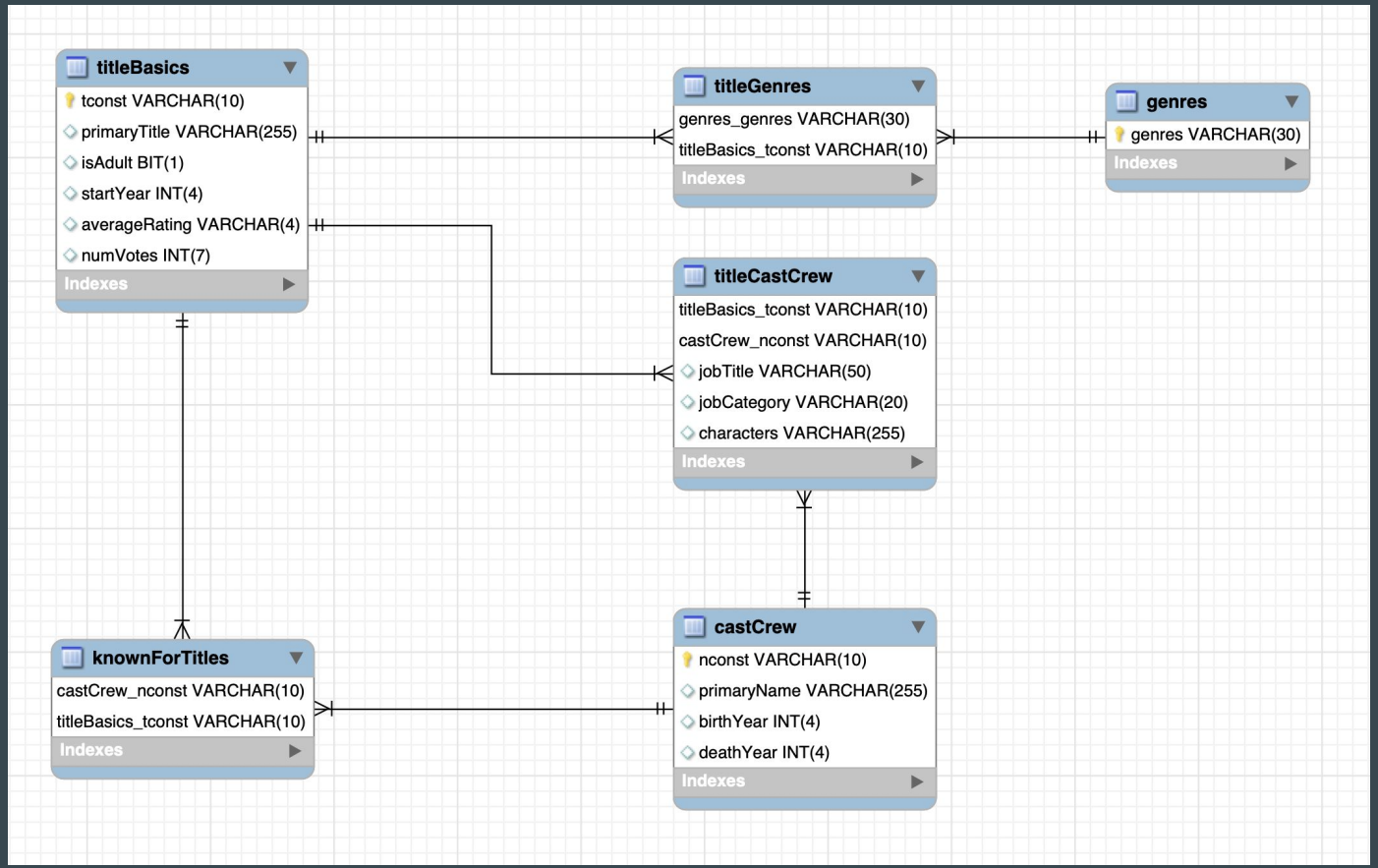
Popularity: we will choose to invest in films and shows that are most likely to be binge watched by our existing members

Rating: we will determine which films and shows are likely to draw praise from online rating sites and critics in order to bring in new subscribers

Tools and Methodologies

- R and Python for data cleansing and transformation
- Google Cloud SQL (MySQL) for Storage
- Tableau for data visualizations

Entity Relationship Diagram



Which 3 people that have worked on more than 1 movie have the highest average film rating with at least 1,000 votes in total?

```
SELECT
    primaryName,
    averageRating,
    totalVotes,
    numMovies
FROM(SELECT
        ROUND(AVG(tb.averageRating),2) AS averageRating,
        SUM(tb.numVotes) AS totalVotes,
        cc.nconst,
        COUNT(tb.tconst) AS numMovies,
        cc.primaryName AS primaryName
    FROM
    titleBasics tb
        INNER JOIN
    titleCastCrew tcc ON tb.tconst = tcc.titleBasics_tconst
        INNER JOIN
    castCrew cc ON tcc.castCrew_nconst = cc.nconst
    GROUP BY
        cc.nconst) AS a
WHERE
    totalVotes >=1000 AND
    numMovies > 1
ORDER BY
    2 DESC
LIMIT
    3;
```

primaryName	averageRating	totalVotes	numMovies
Michael Rapaport	8.05	1828	2
Frank Welker	7.35	1444	2
Vincent Price	6.73	9308	3

Do adult films have a higher or lower rating and number of views than non-adult films?

```sql
SELECT
    'Not Adult' AS isAdult,
    ROUND(AVG(averageRating),2) AS averageRating,
    SUM(numVotes) AS totalVotes,
    COUNT(isAdult) AS countMovies
FROM
    titleBasics
WHERE isAdult = 0
UNION
SELECT
    'Adult' AS isAdult,
    ROUND(AVG(averageRating),2) AS averageRating,
    SUM(numVotes) AS totalVotes,
    COUNT(isAdult) AS countMovies
FROM
    titleBasics
WHERE isAdult = 1;
```

isAdult	averageRating	totalVotes
Not Adult	6.88	6271483
Adult	6.42	2635

What genres have the highest ratings on average?

```
SELECT
    g.genres,
    ROUND(AVG(tb.averageRating),2) AS averageRating,
    ROUND(AVG(tb.numVotes),0) AS averageNumVotes
FROM
titleBasics tb
    INNER JOIN
titleGenres tg ON tb.tconst = tg.titleBasics_tconst
    INNER JOIN
genres g ON tg.genres_genres = g.genres
WHERE g.genres != 'NA'
GROUP BY 1
ORDER BY 2 DESC;
```

genres	averageRating	averageNumVotes
History	7.59	64
Western	7.33	48
Sport	7.28	21
Documentary	7.14	78
War	7.12	97
Biography	7.11	947
Crime	7.09	535
Adventure	7.05	2323
Musical	6.98	40
Drama	6.92	905
Music	6.92	41
Comedy	6.89	998
Animation	6.89	169
Action	6.81	2867
Short	6.79	28
Adult	6.49	16
Family	6.49	100
Romance	6.41	81
Fantasy	6.40	2059
Mystery	6.39	170
Thriller	6.04	148
Horror	5.71	1306
News	4.86	100

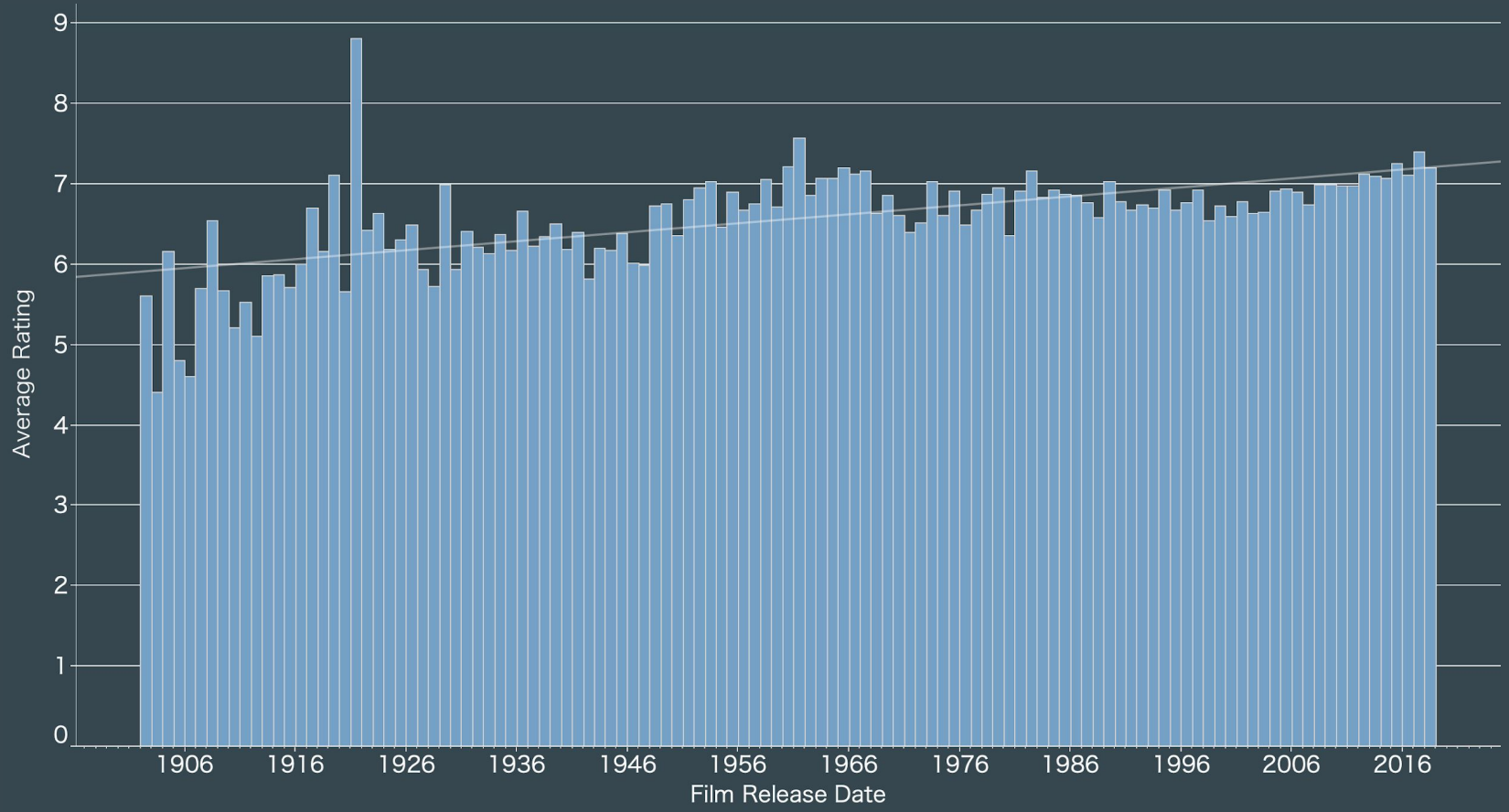
What are the types of content have been produced over the decades?

```sql
SELECT
    g.genres,
    COUNT(IF(startYear < 1920, 1, NULL)) 'Pre 1920s',
    COUNT(IF(startYear BETWEEN 1920 AND 1930, 1, NULL)) '1920s',
    COUNT(IF(startYear BETWEEN 1930 AND 1950, 1, NULL)) '1930s',
    COUNT(IF(startYear BETWEEN 1940 AND 1950, 1, NULL)) '1940s',
    COUNT(IF(startYear BETWEEN 1950 AND 1960, 1, NULL)) '1950s',
    COUNT(IF(startYear BETWEEN 1960 AND 1970, 1, NULL)) '1960s',
    COUNT(IF(startYear BETWEEN 1970 AND 1980, 1, NULL)) '1970s',
    COUNT(IF(startYear BETWEEN 1980 AND 1990, 1, NULL)) '1980s',
    COUNT(IF(startYear BETWEEN 1990 AND 2000, 1, NULL)) '1990s',
    COUNT(IF(startYear BETWEEN 2000 AND 2010, 1, NULL)) '2000s',
    COUNT(IF(startYear >= 2010, 1, NULL)) '2010s',
    COUNT(*) AS Total
FROM
titleBasics tb
    INNER JOIN
titleGenres tg ON tb.tconst = tg.titleBasics_tconst
    INNER JOIN
genres g ON tg.genres_genres = g.genres
WHERE g.genres != 'NA'
GROUP BY 1;
```

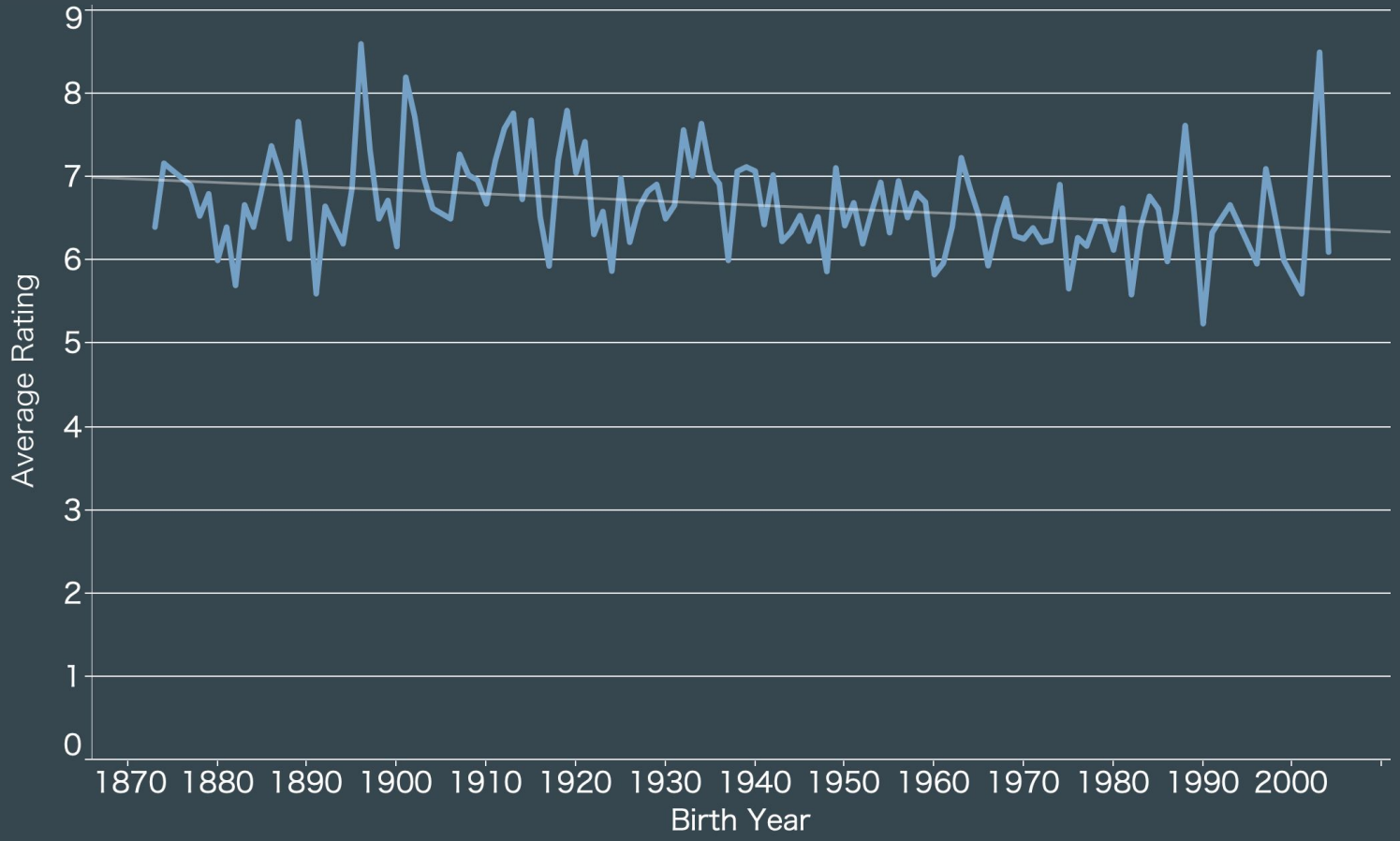
genres	Pre 1920s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s	Total
Action	1	2	18	13	15	42	44	99	129	235	242	771
Adult	0	0	0	0	0	0	20	29	49	41	12	140
Adventure	2	1	6	4	9	22	34	38	42	71	82	288
Animation	3	9	29	14	20	28	28	32	59	96	118	384
Biography	1	1	1	0	0	2	3	7	22	48	47	118
Comedy	16	12	51	30	51	114	113	141	228	461	560	1614
Crime	1	1	17	7	10	17	44	51	64	106	123	385
Documentary	15	4	11	7	9	16	18	37	71	237	208	584
Drama	21	12	58	28	44	62	84	124	162	351	372	1191
Family	0	0	2	0	2	2	8	2	12	32	21	74
Fantasy	1	1	2	1	1	2	1	4	2	9	6	27
History	0	0	1	1	0	0	1	1	0	4	7	14
Horror	0	0	1	0	2	2	11	5	13	40	62	121
Music	0	0	4	3	1	0	6	11	19	37	36	102
Musical	1	0	6	5	1	1	1	1	1	3	3	17
Mystery	1	1	4	1	1	0	3	0	5	4	8	26
News	1	0	0	0	0	0	0	0	4	5	2	12
Romance	3	0	5	3	1	5	2	2	3	10	12	40
Short	8	2	6	3	5	8	10	11	38	66	33	175
Sport	0	0	1	1	0	2	2	5	8	14	14	38
Thriller	0	0	0	0	1	1	0	4	13	17	18	49
War	0	0	0	0	0	1	2	1	2	1	0	6
Western	0	2	12	7	27	30	5	1	1	1	1	71

Tableau Visualizations

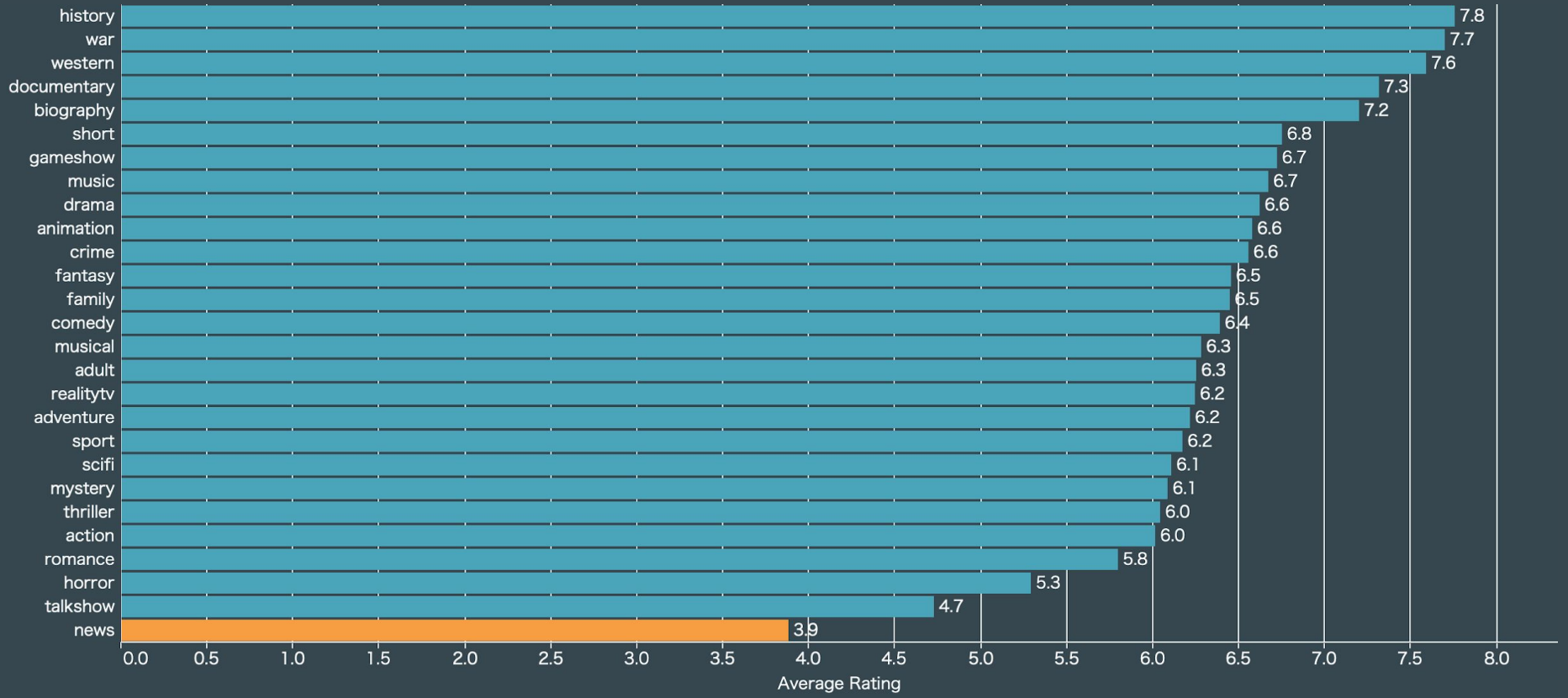
Film Ratings by Release Year are Increasing...



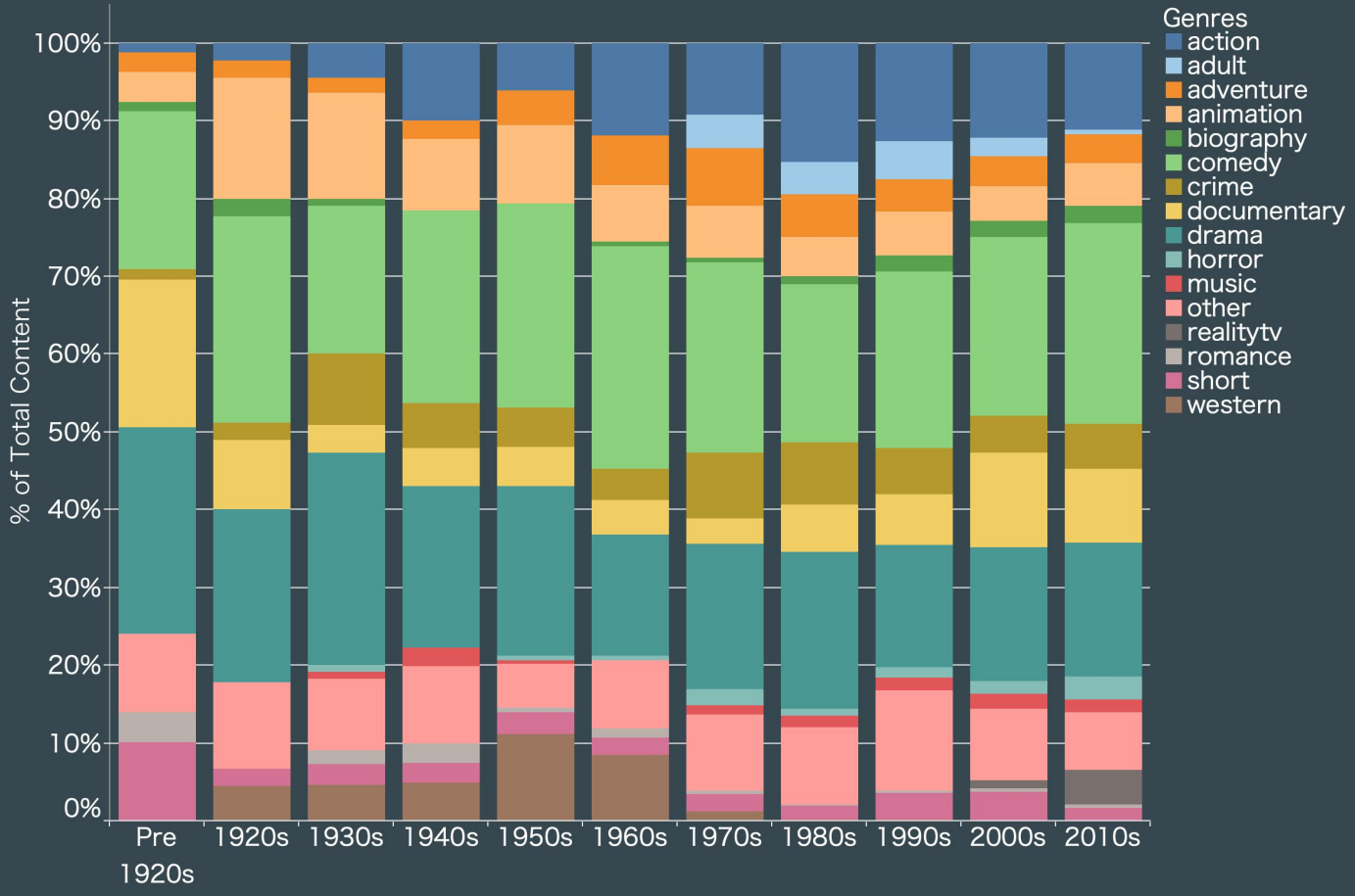
...While Film Ratings by Cast Birth Year are Decreasing



Genres by Rating: Fake News?

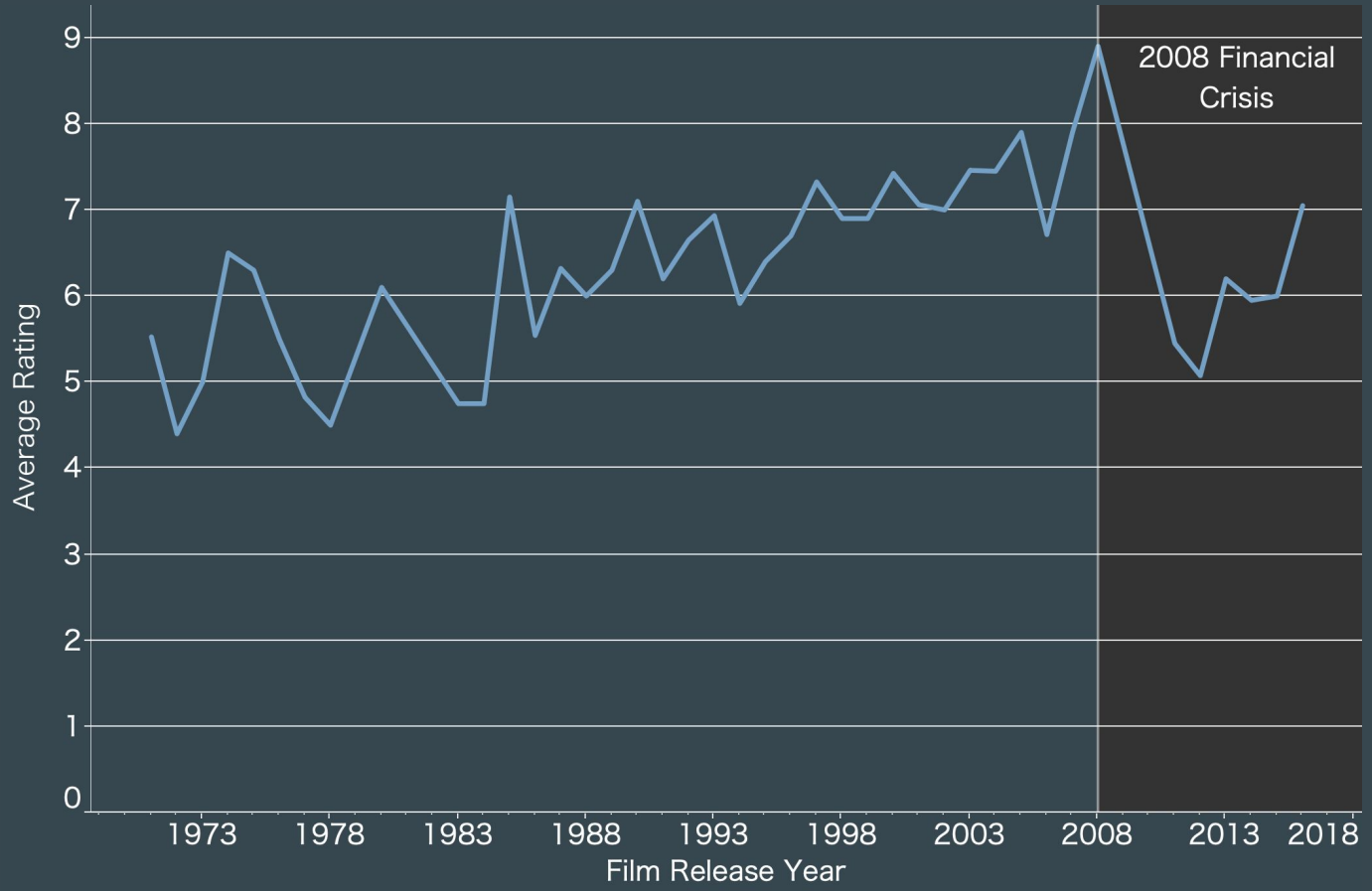


Genres by Decade



- Western Films peaked during the 1950s
- Reality TV came into view in the new millenium
- Action films steadily increased, peaking in the 1980s
- Adult films were most popular during the end of the 20th century

Adult Film Ratings Fell During the Recession

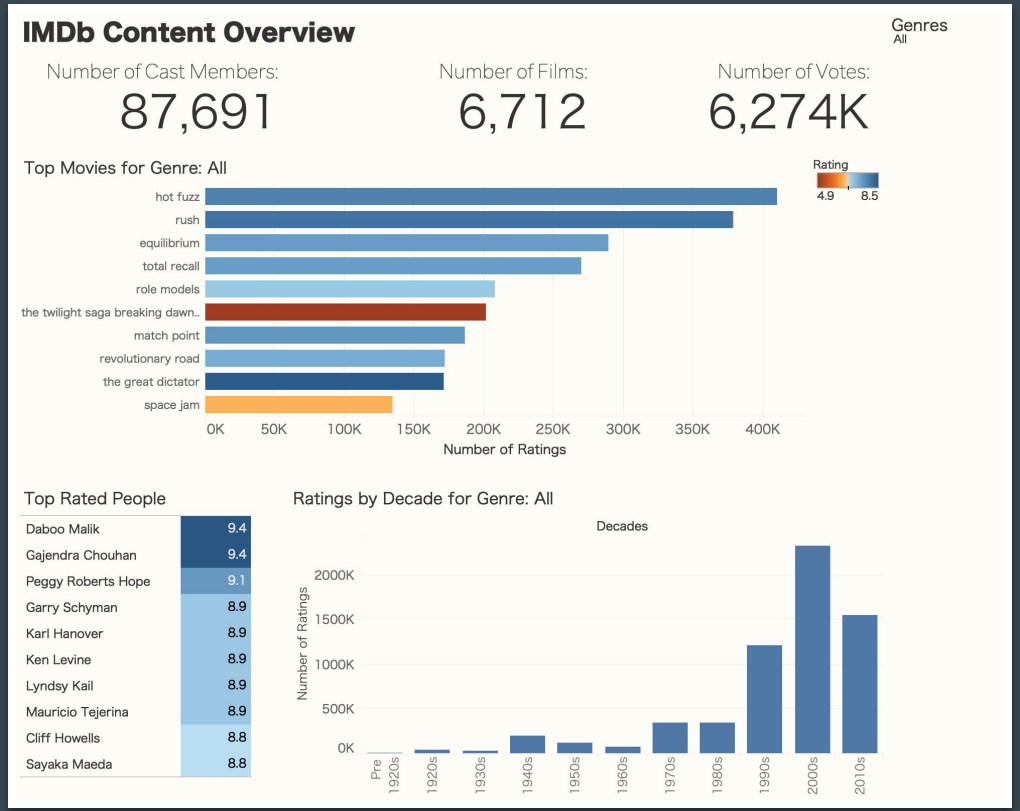


Most Common Used Words in Content Titles

Interactive Dashboard

Link to Dashboard on Tableau Public:

https://public.tableau.com/profile/jamie.olds#!/vizhome/FinalProjectDashboard_8/ContentOverview



Recommendations

Benefit from getting data on:

- Box office revenues
- Ratings from other sites (e.g. Rotten Tomatoes, MetaCritic)
- Twitter and Facebook followings and sentiment (proxy for popularity)
- Book sales/ratings if adapted from books

Scope for improvement:

- Have a filter or attribute for different languages
- Have an attribute for “known for titles” to remove circular reference
- Use Neo4j, MongoDB, or other NoSQL databases to better understand the data

Lessons Learned

- Data size limitations
  - SQL queries timing out or taking hours
  - Content not being correctly matched with the cast, and vice versa
- Handling data in different languages which use non-latin characters
- Capabilities of using Google Cloud SQL, speed and transfer rate limitations
- Dealing with NSFW data