IMDb Movie Analysis

ECE 180 Team 7

Aiswarya Akumalla/Liyan Chen/Qiuhao Cao/Sean Kamano/Xiaoguang Wang/Xinhe Zhang/Zhisheng Huang

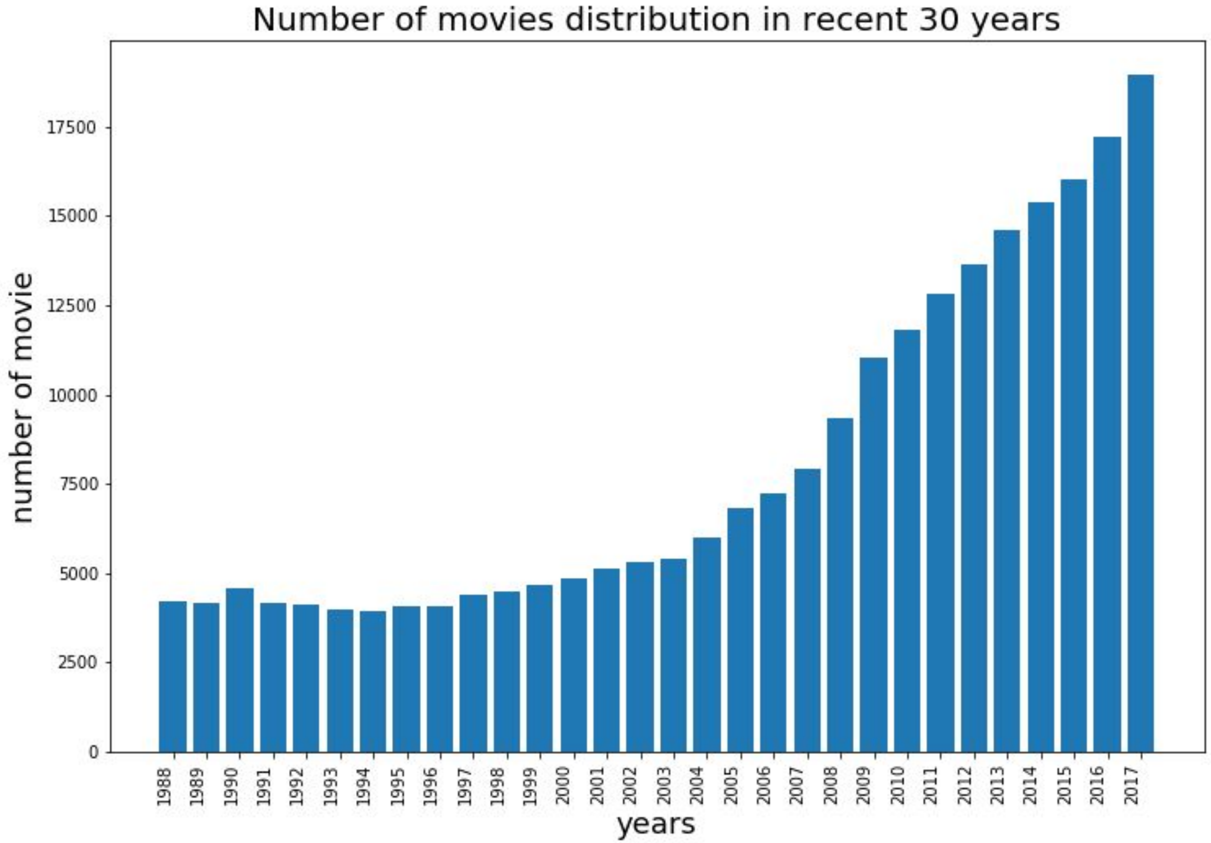
Outline

- Motivation
- Dataset
- Data Cleaning
- Results
    - Rating
    - Time
    - Location
    - Connection Network
- Conclusion

Motivation

- Analyze the IMDb dataset to provide information for distributors, movie production firms and marketing firms
- Understand aspects of this industry such as:
    - The composition of all movies in the dataset on based on certain properties (genre, region, release date, etc.)
    - Professional networks and popularity of actors through visualization

Distribution of movies over time



Dataset

- IMDb (Internet Movie Database)
  - 478,145 movies
  - basics: genre, release year, and runtime
  - crew: directors and writers
  - principals: top-billed cast/crew
  - ratings: number of votes and weighted average rating
  - name: information about cast/crew (associated movies, birth and death year, and profession)
  - region: ISO-A2 (two-character) country codes for the country each movie was released in
- TMDb (The Movie Database)
  - person: list of popular actor along with popularity scores

Data Cleaning

- Filtered titles by type “Movie”
- Dropped rows with missing values
- Dropped rows where number of Votes for a movie < 100,000
- Split values of a column and duplicate corresponding rows

tconst	directors	avgRating	# of Votes
tt0000233	nm0000690	8.4	134500
tt0000234	“nm0000400, nm0000500, nm0000600”	7.4	107800

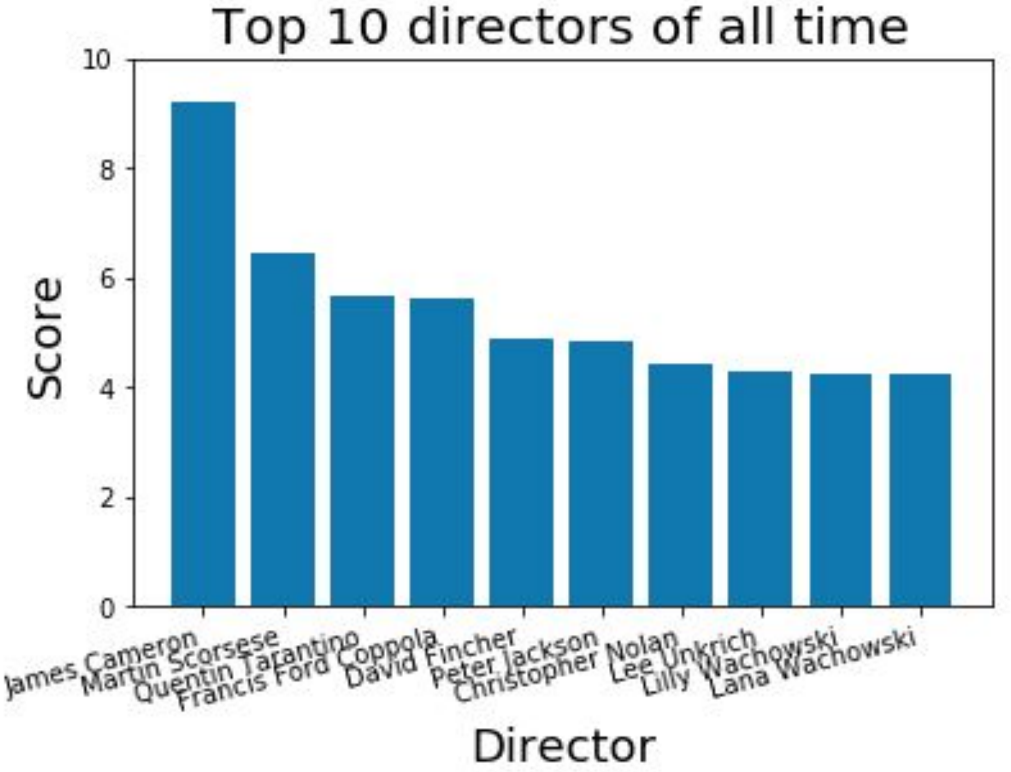
tconst	directors	avgRating	# of Votes
tt0000233	nm0000690	8.4	134500
tt0000234	nm0000400	7.4	107800
tt0000234	nm0000500	7.4	107800
tt0000234	nm0000600	7.4	107800

Results -- Top 10 directors

- Average Rating for a title = R
- Number of votes for a title = V
- # of movies made by director threshold > 4

Score for a director = R * V / no. of movies

tconst	directors	avgRating	numVotes
tt0040746	nm00000033	8.0	106921
tt0044079	nm00000033	8.0	107709
tt0055600	nm00000033	8.2	361351
tt0045906	nm00000033	8.5	124765



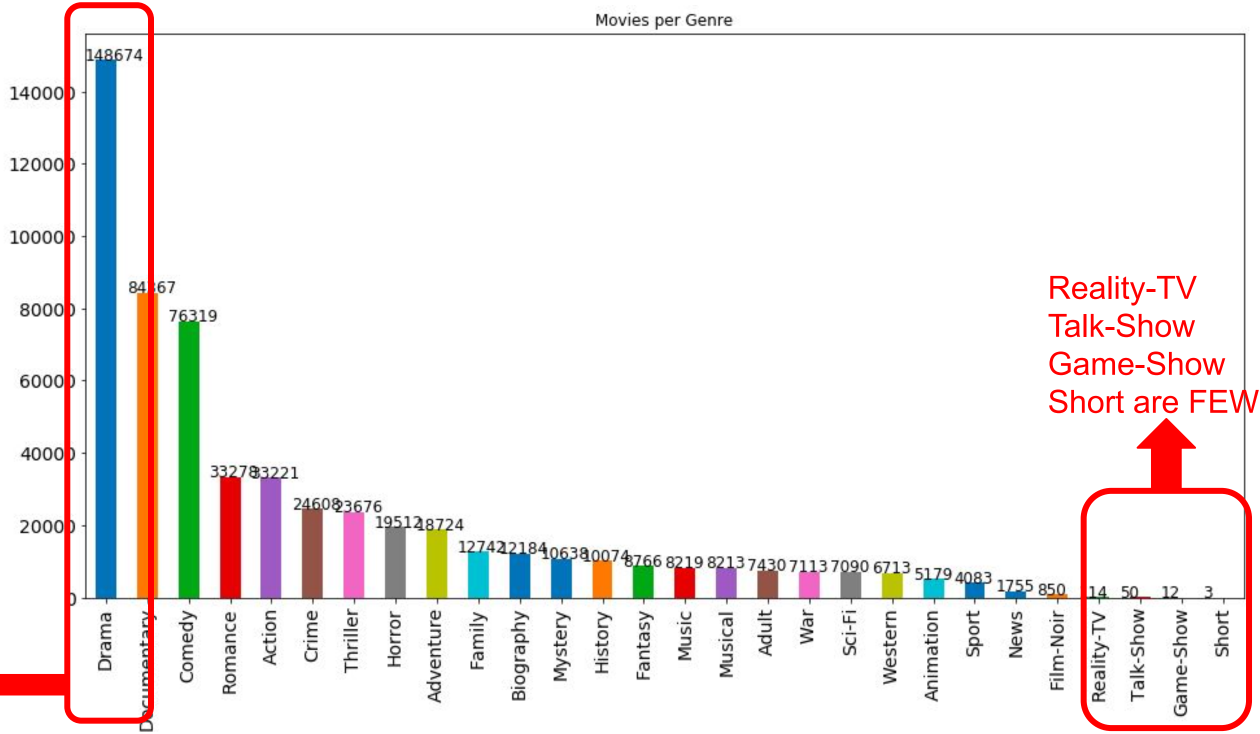
James Cameron
Martin Scorsese
Quentin Tarantino
Francis Ford Coppola
David Fincher
Peter Jackson
Christopher Nolan
Lee Unkrich
Lilly Wachowski
Lana Wachowski

Results -- Distribution of genres

- 364461 movies from 1894 to 2018
- Multiple genre titles:

Ex: Movie is categorized as Drama/Comedy

Drama count increased by 1
Comedy count increased by 1

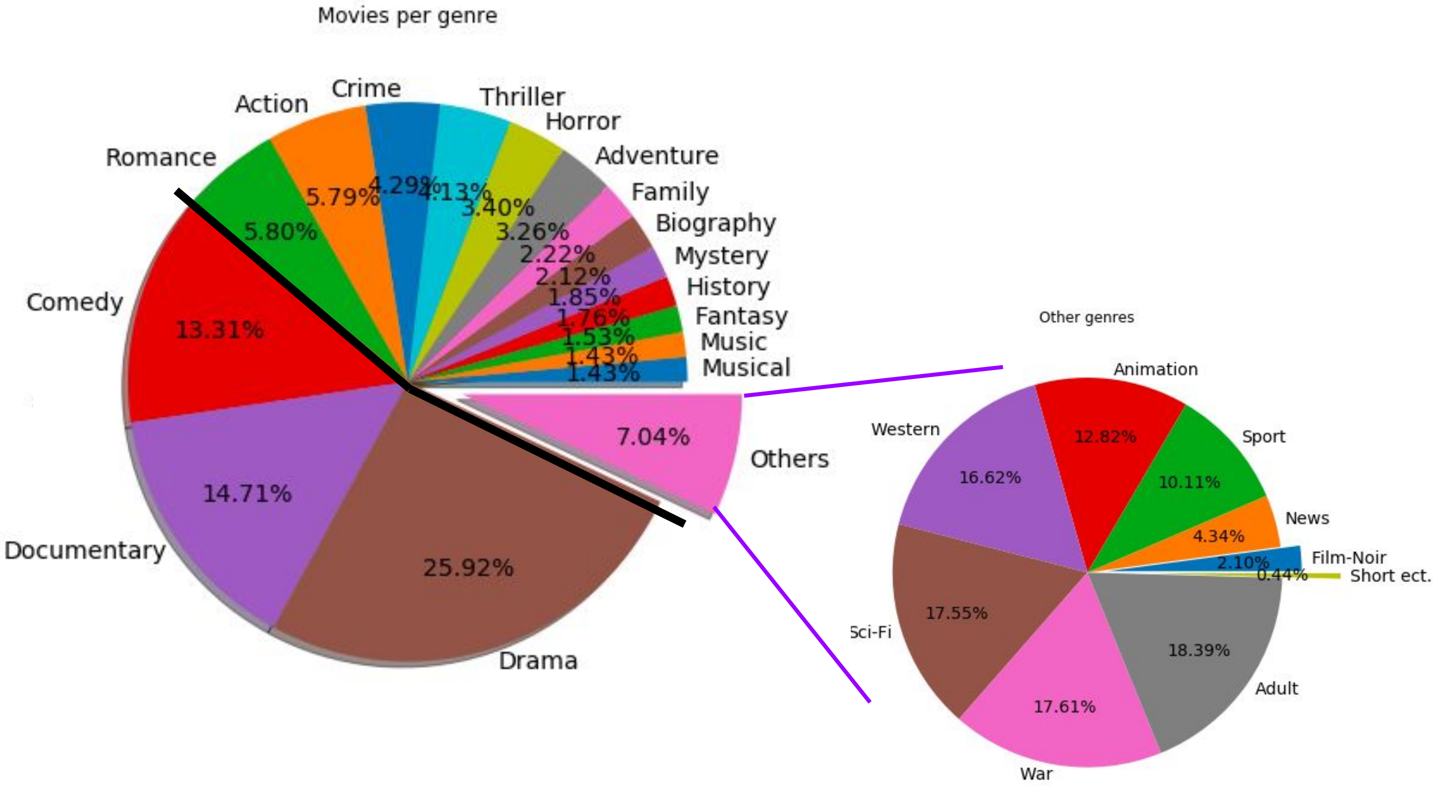


People love **Dramas!**

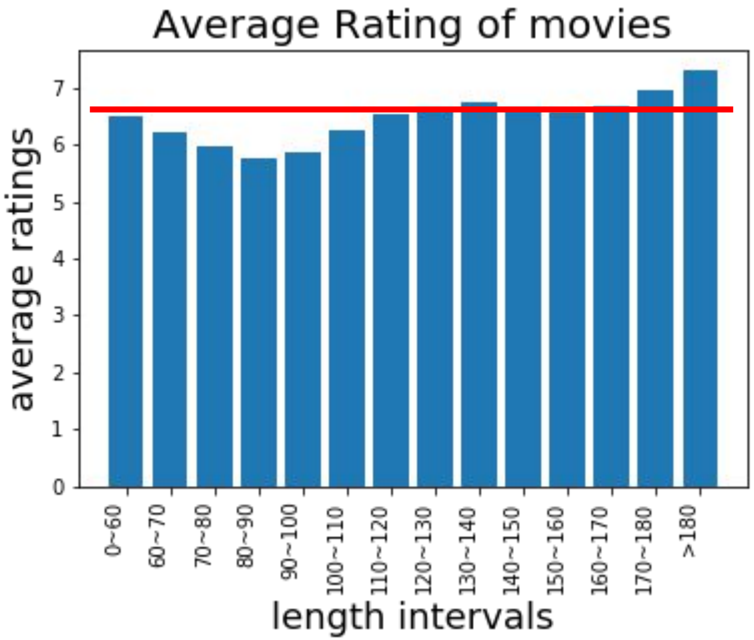
Reality-TV
Talk-Show
Game-Show
Short are FEW.

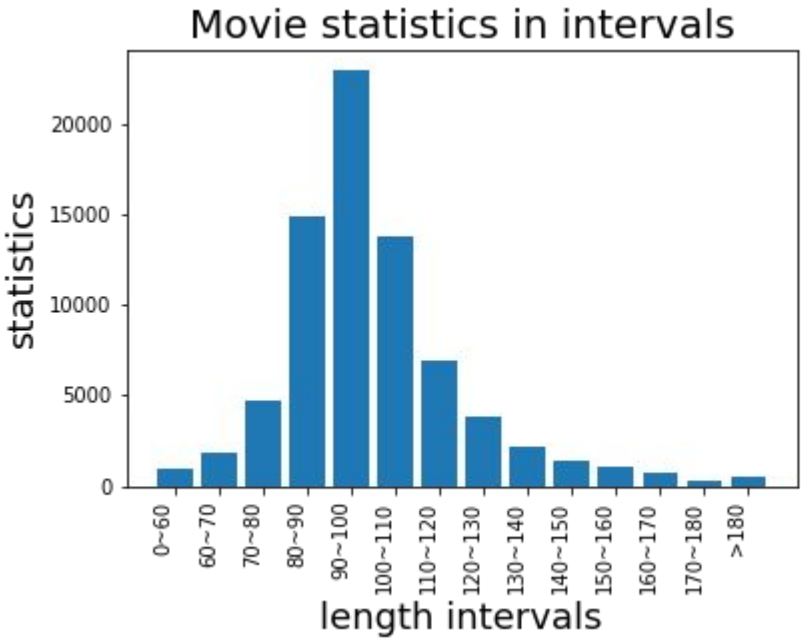
Results -- Distribution of genres(Cont.)

- Half of movies are **comedy**, **documentary** and **drama**.

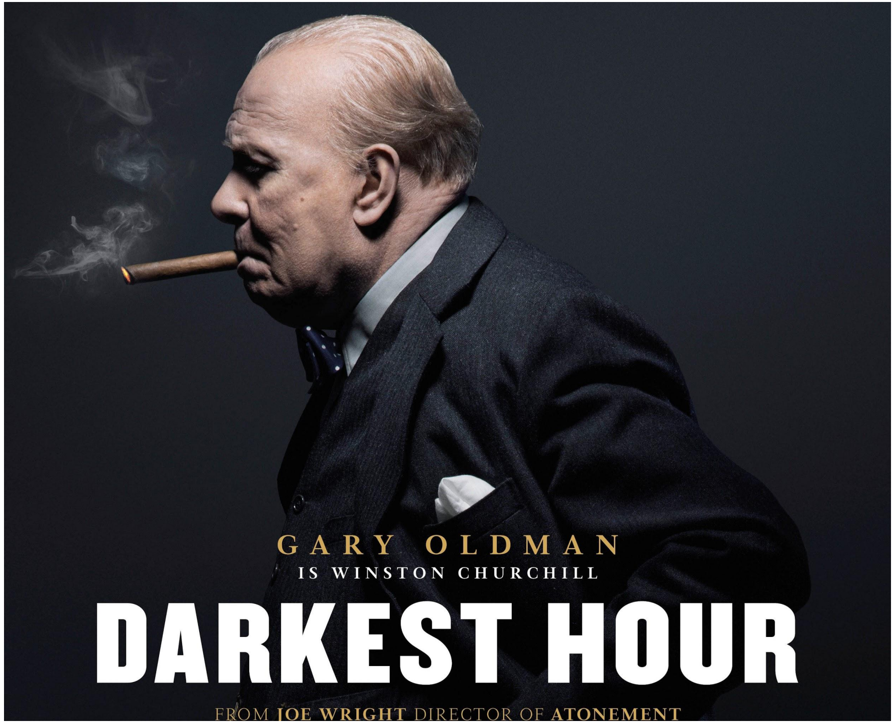


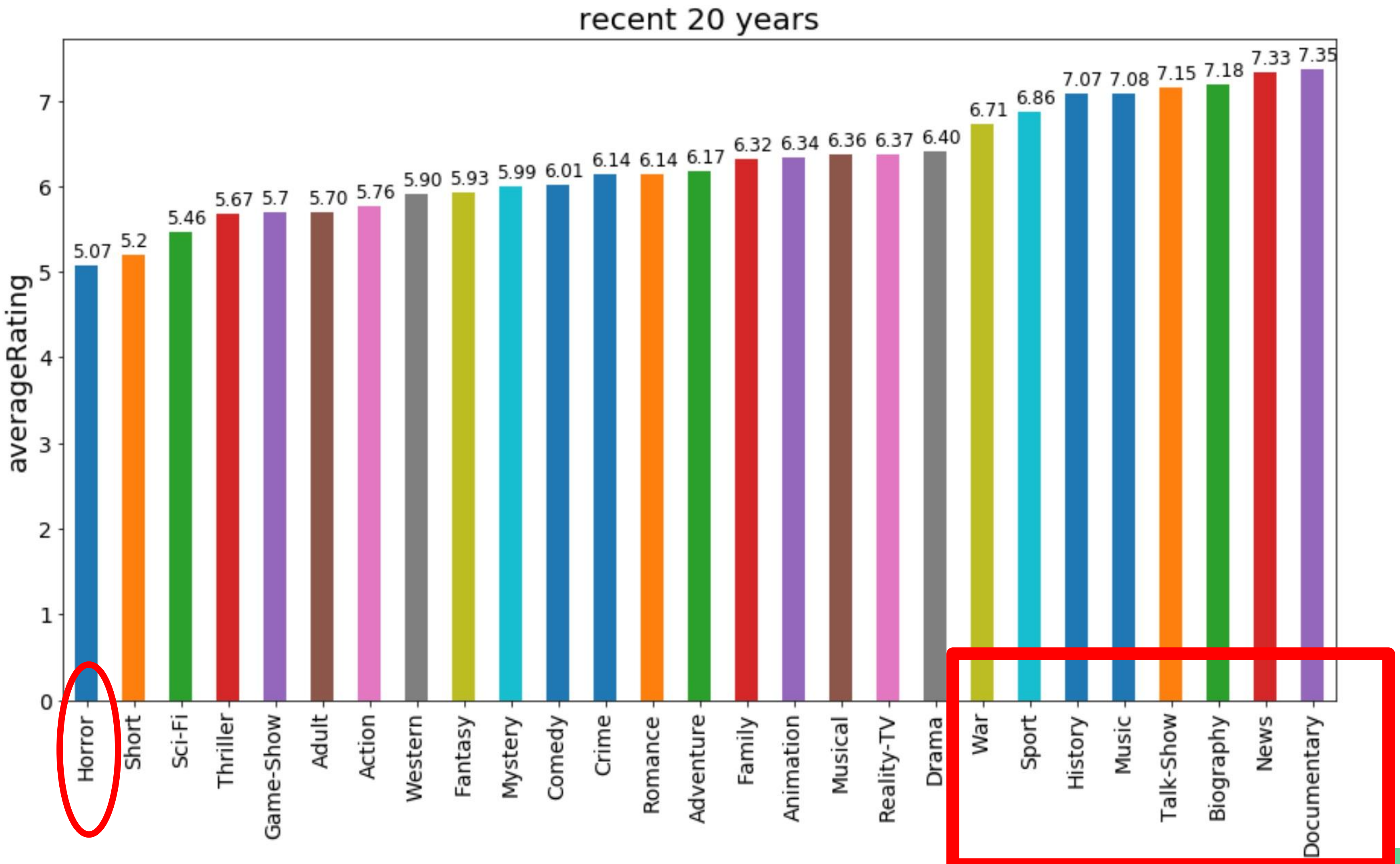
Results -- Movie length and ratings


Average Rating of movies

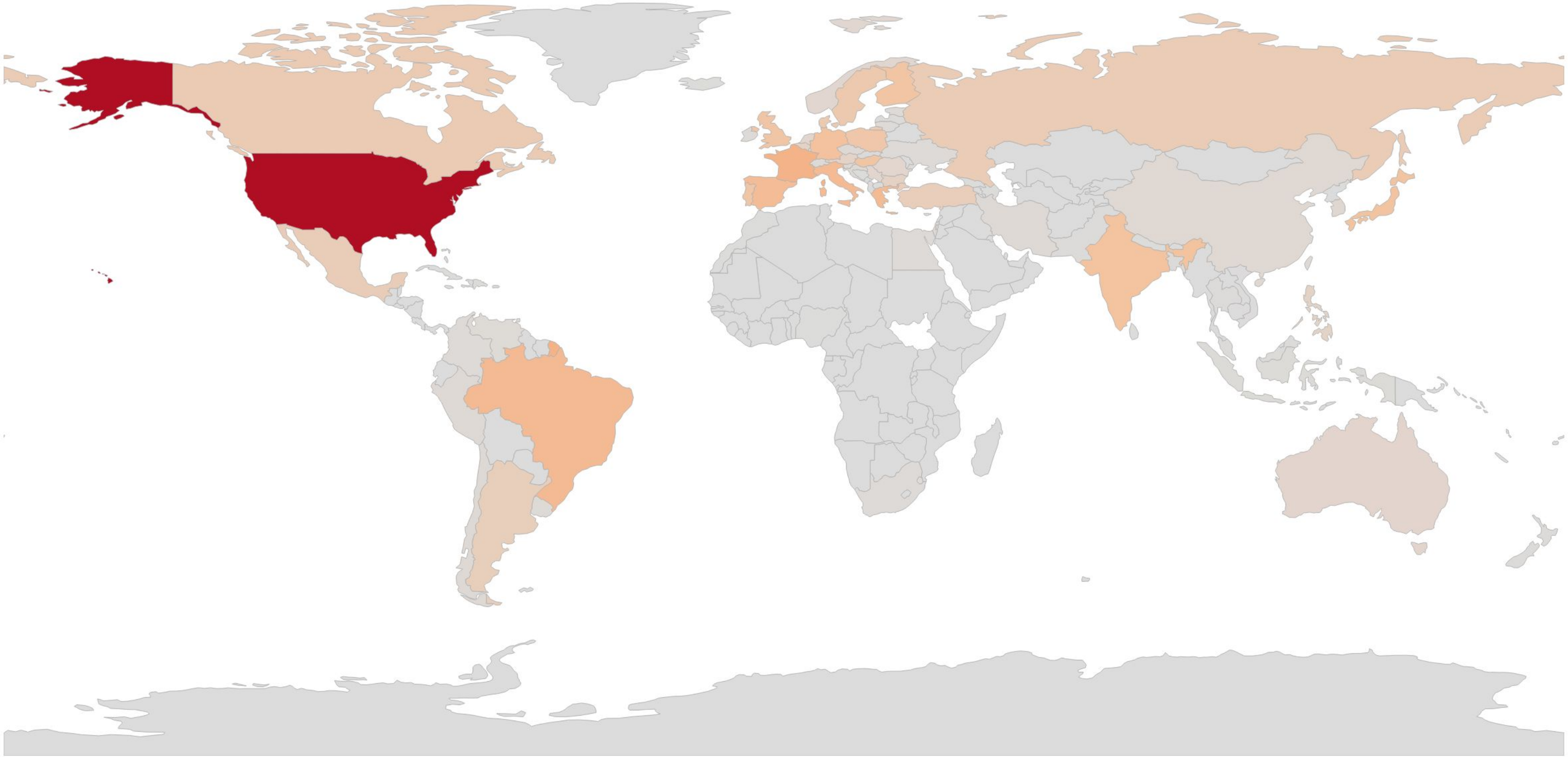

Movie statistics in intervals

Results -- Genres and ratings




recent 20 years
averageRating
0
1
2
3
4
5
6
7
5.07
5.2
5.46
5.67
5.7
5.70
5.76
5.90
5.93
5.99
6.01
6.14
6.14
6.17
6.32
6.34
6.36
6.37
6.40
6.71
6.86
7.07
7.08
7.15
7.18
7.33
7.35
Horror
Short
Sci-Fi
Thriller
Game-Show
Adult
Action
Western
Fantasy
Mystery
Comedy
Crime
Romance
Adventure
Family
Animation
Musical
Reality-TV
Drama
War
Sport
History
Music
Talk-Show
Biography
News
Documentary

Results -- Distribution of movies among countries

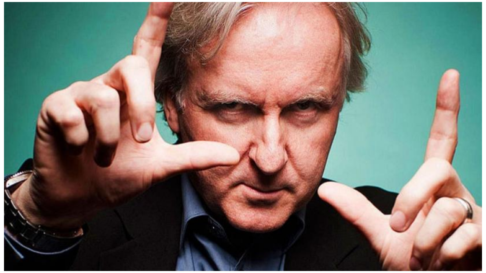


Results -- The connection network

- Connections between the 100 most popular actors and actresses (which movies they have starred in together)
- Shows how popular they are relative to each other based on a popularity score
  - Popularity score: 0-20, log(number of votes * average rating)

Concluding Remarks

- Significant growth in the output of movies over the last decade
- Analysis of ratings reveals information about the movie industry such as:
  - Most well-received directors
  - Consumer preference for dramas, comedies, and documentaries
  - Largest audiences in the US and the West
  - Relationships between actors/actresses and their popularity



+

+

+ Drama/Documentary
U.S.

Future Work

- Through the information gained from our project, we can figure out the most important factors which affect ratings

  Example: ratings-based recommender system

- What matters:
  - Genre, directors, actors/actress
- What doesn’t:
  - Movie length

Thank You

Happy movie watching!