NETFLIX

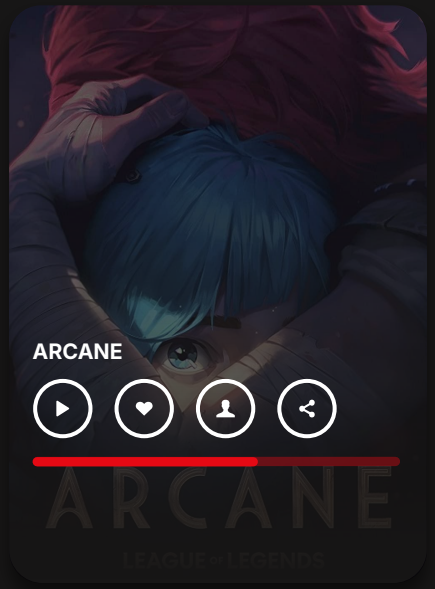
DATA PROJECT

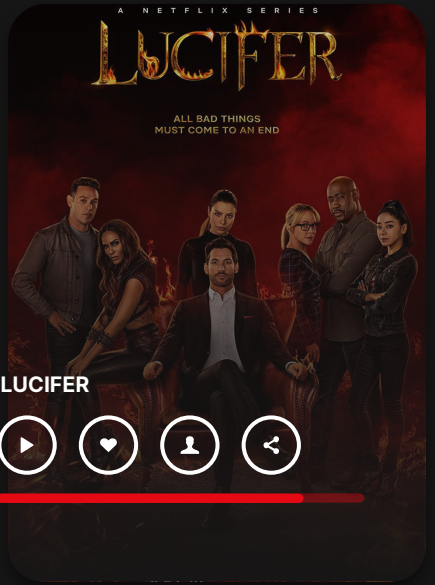
BACKGROUND & MOTIVATION

We aim to explore Netflix's dataset of shows and movies to gain insights into viewing habits, popular genres, and content distribution across countries. This dataset spans from 1925 to 2021 and includes various entries and variables such as titles, descriptions, imdb ratings, and categories. This will help us understand the relationship between the content features and their popularity on Netflix. The resulting data will be useful for Netflix to make data-driven decisions about content acquisition and recommendations, helping cater better to global user preferences. We will focus on text mining techniques to analyze descriptions and genre tags and conduct a study to identify key factors linked with the popularity of shows and movies. This project can assist in predicting future content success.

WHAT?


NARCOS
RISE OF THE CARTELS
NARCOS

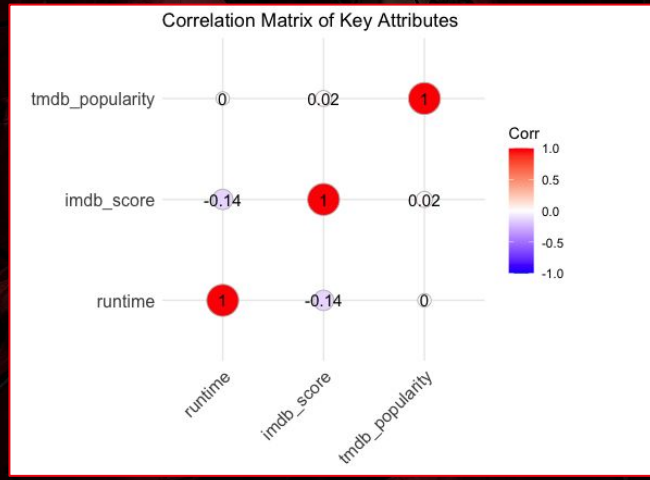

ARCANE
LEAGUE of LEGENDS
ARCANE


A NETFLIX SERIES
LUCIFER
ALL BAD THINGS
MUST COME TO AN END
LUCIFER

What factor—runtime, IMDb score, or TMDb popularity—correlates most strongly with the popularity of shows and movies over time?

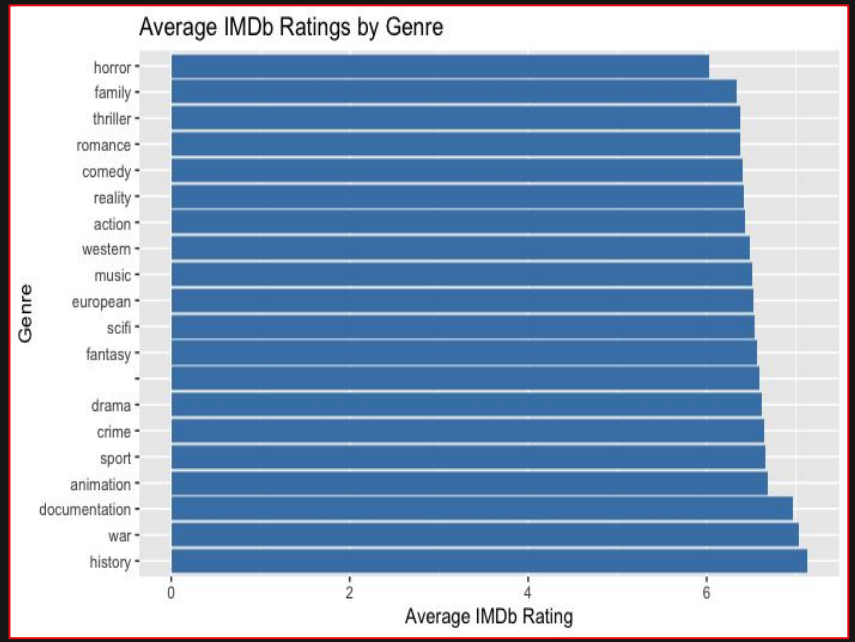
1. **Runtime:** The duration of a show or movie.
2. **IMDb Score:** A measure of content quality based on ratings.
3. **TMDb Popularity:** A metric reflecting how often content is searched or viewed.

- Runtime and IMDb Score: Weak negative correlation (-0.14), suggesting runtime has minimal impact on IMDb scores.
- TMDb Popularity and IMDb Score: Very weak positive correlation (0.02), indicating little to no direct relationship between popularity and ratings.
- Runtime and TMDb Popularity: No significant correlation (0).
- Runtime, while often considered important, does not strongly influence IMDb scores or popularity.
- TMDb popularity is largely independent of runtime or IMDb scores, suggesting other factors (e.g., marketing or trending content) might drive popularity.



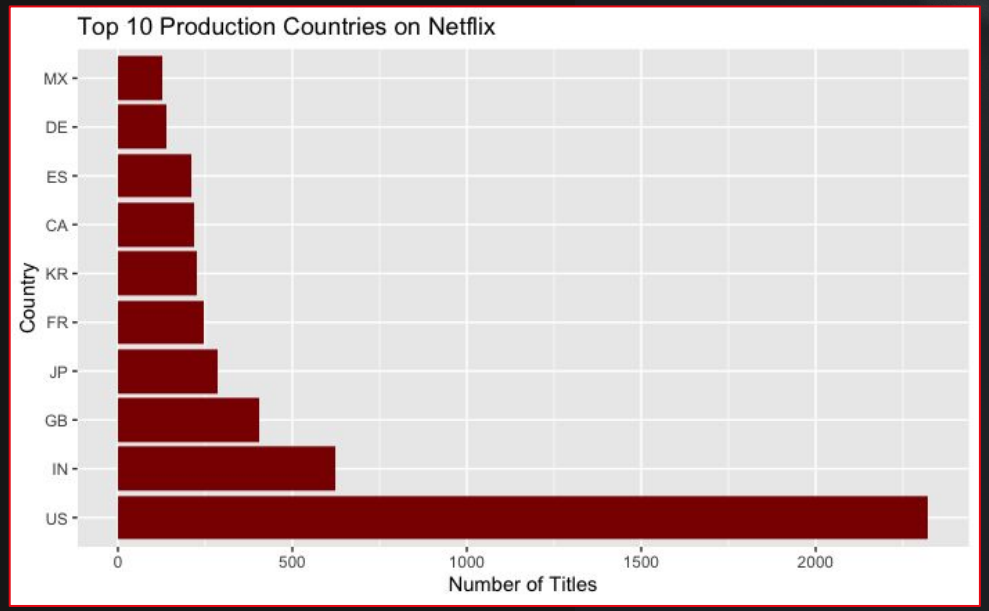
What Genres Are Associated with Higher IMDb Ratings?

- Horror, Family, and Thriller genres have the highest average IMDb ratings.
- History and War genres show relatively lower ratings, indicating niche or less mainstream appeal.
- The insights can help in targeting popular genres for new content



Average IMDb Ratings by Genre

What are the most prolific production countries?



Top Production Countries on Netflix

- United States: Dominates Netflix's content library with over 2,000 titles, showcasing its pivotal role in global content production.
- India: Second most prolific, reflecting its growing importance in Netflix's content strategy.
- United Kingdom: Third, contributing significantly to Netflix’s international appeal.
- Countries like Japan, France, and South Korea indicate Netflix’s efforts to diversify its content geographically.
- European markets (e.g., Germany, Spain, France) collectively contribute a significant number of titles.
- Netflix’s reliance on U.S.-based content aligns with global demand but may lead to overrepresentation of U.S. perspectives.
- The rise of India, Japan, and South Korea suggests growing popularity of Asian content, aligning with Netflix's push into new markets.

HOW

How do IMDb scores vary for movies vs. TV shows?

Median Scores: Movies have a slightly higher median IMDb score than TV shows.

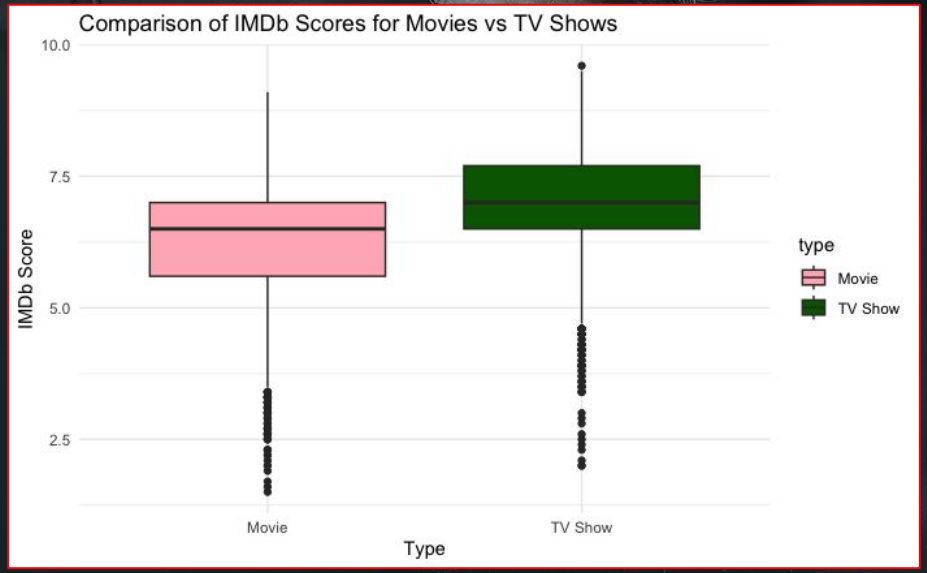
Score Distribution: TV shows exhibit a wider range of scores, with more variability compared to movies.

Insights:

- Movies: Generally maintain consistent ratings, indicating higher and more uniform production quality.
- TV Shows: The variability suggests diverse audience preferences and potentially inconsistent quality.

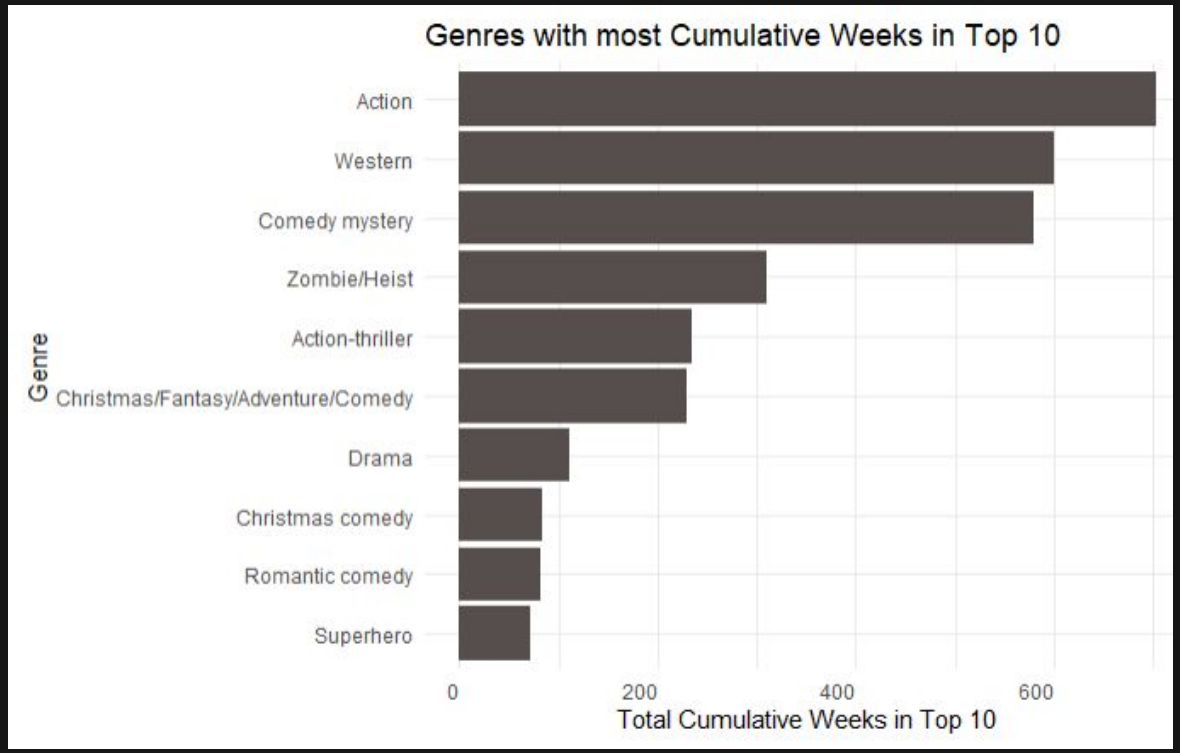
Strategic Recommendations for Netflix:

- Focus on improving TV show quality to reduce the lower-rated outliers.
- Leverage the consistency in movie ratings by investing in more high-quality movie productions.



Comparison of IMDb Scores: Movies vs. TV Shows

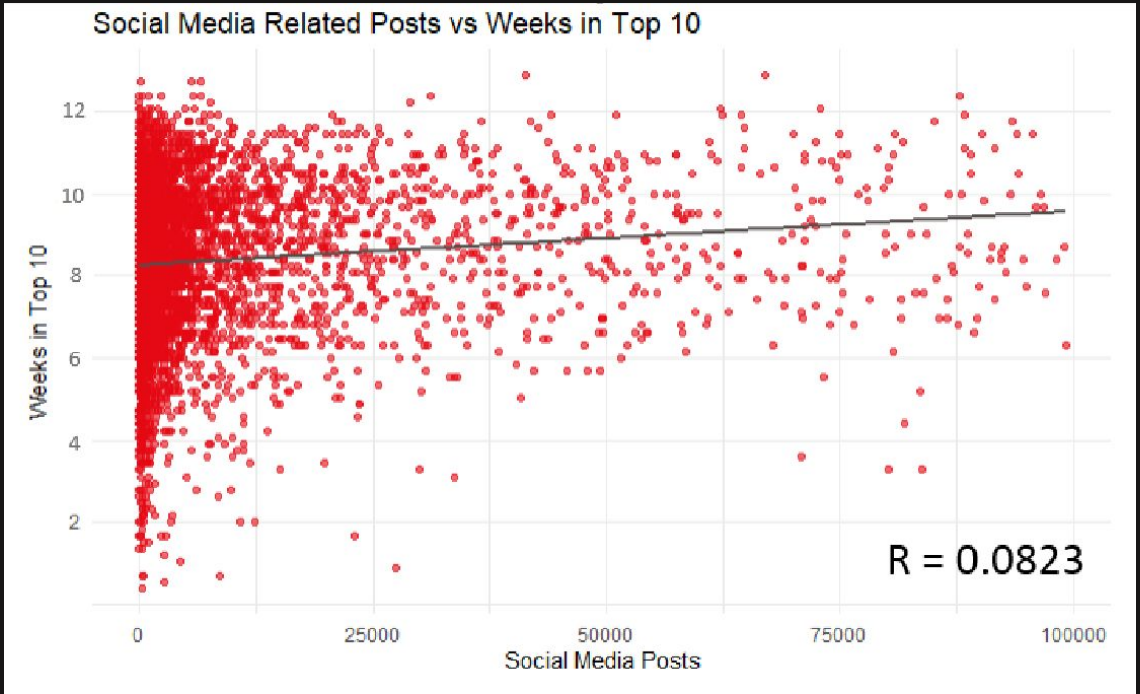
Top Genres by Cumulative Weeks



Question 5: How do social trends influence the relevance across content of various types?

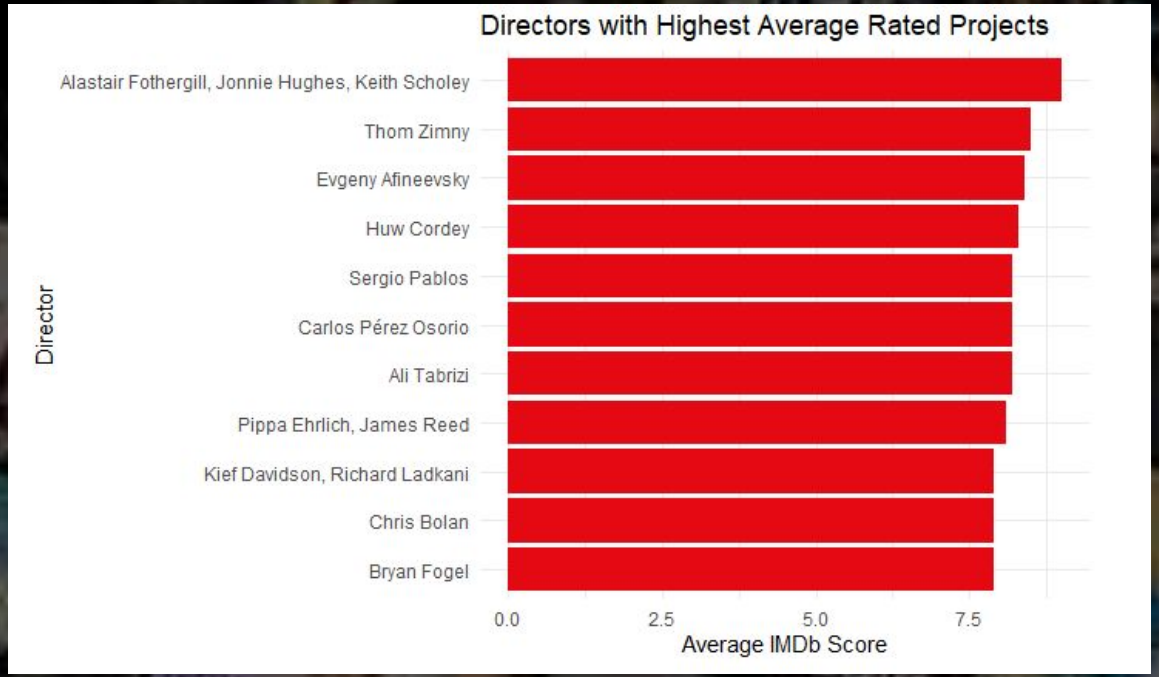
- Action, Western, and Comedy mystery genre shows have the most cumulative weeks in the top 10 charts
- Romantic Comedy and Superhero titles contain the least cumulative weeks

Social Media Posts vs Cumulative Weeks in Top 10



- There is a weak positive correlation between social media posts related to a title and the total amount of weeks it made in the top 10 charts
- Increased social media activity shows a slight link to a title longevity in the top 10 charts,
- Small but noticeable

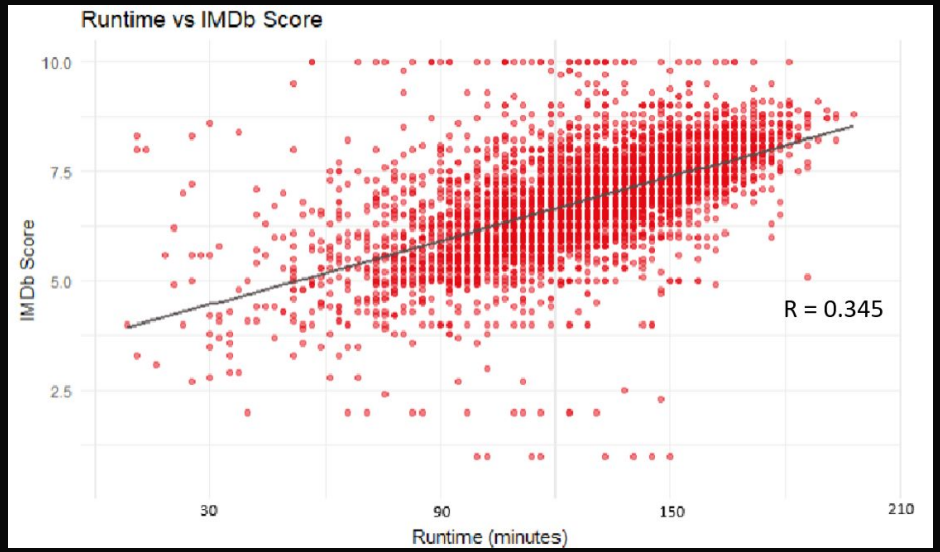
Top Directors with the Highest Rated Projects



Question 6: How do key factors of filmmaking influence customer preferences and ratings for movies and shows?

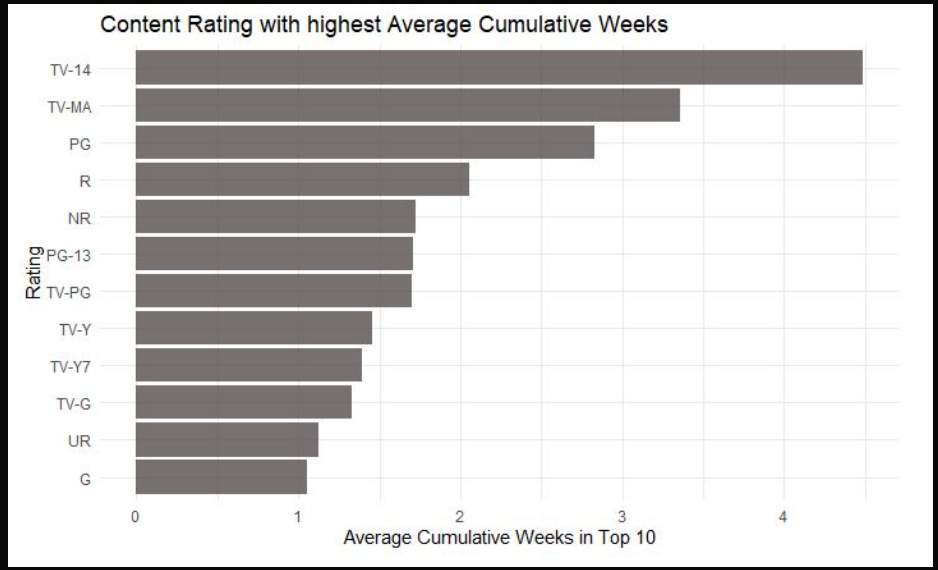
- Alastair Fothergill, Jonnie Hughes, Keith Scholey, Thom Zimny, and Evgeny Afineevsky have the highest rated projects
- Overall, the average IMDb scores is close among the directors listed

Project Runtime vs IMDb Score



- There is a more stronger correlation between the length of a project in minutes and the IMDb Score
- The longer a project is, the higher the ratings are for it
- There are noticeable exceptions

Audience Engagement based on Content Rating



Runtime: The content rating of a show or movie (Ex. TV-MA, PG-13) can play a crucial role in determining how popular it is among audiences

- TV-14 and TV–MA are the most popular (older audiences)
- UR and G are the least popular (Overall)

TEXT MINING

How can Unstructured Data be Used

Types

- Comments/Reviews
- Movies Descriptions
- Transcriptions
- Social Media Information

Assumptions

- Bag of Words
- Words are Independent
- Stemming

Text-Mining with Movie Descriptions

Organize by Genre

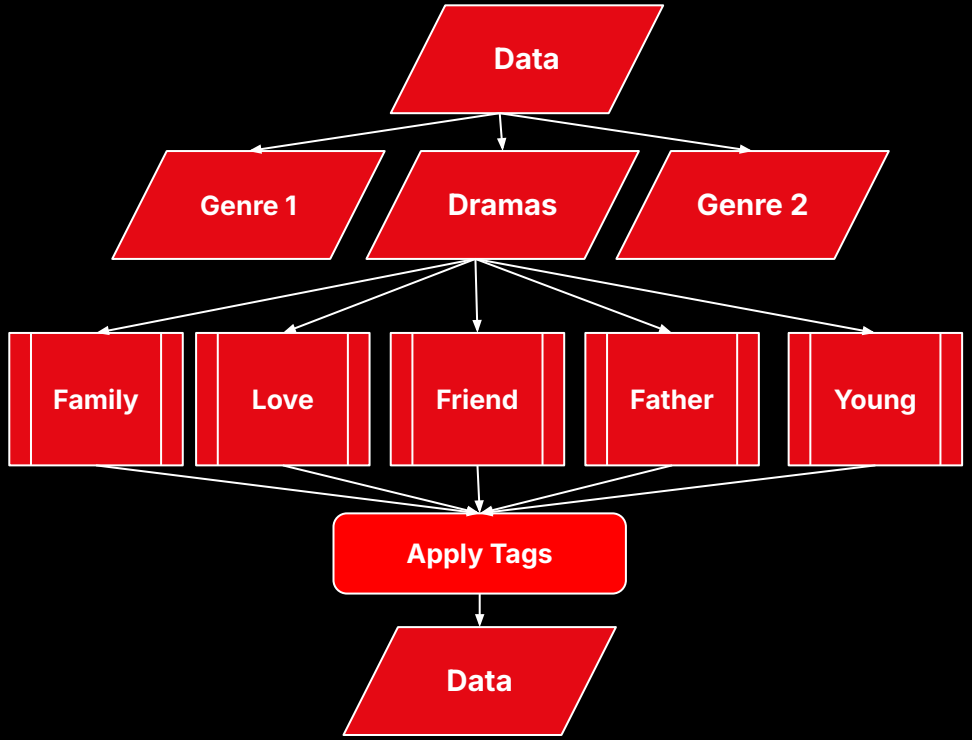
- Separate Genres

Identify Subgenres

- Create a better way to differentiate movies

Tag Subgenres

- Identify movies with key subgenres

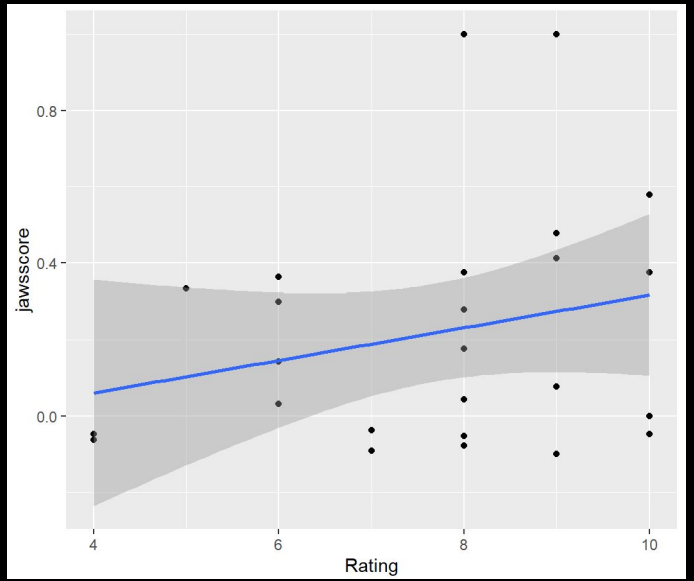


Text-Mining with IMDB Reviews

Sentiment Analysis

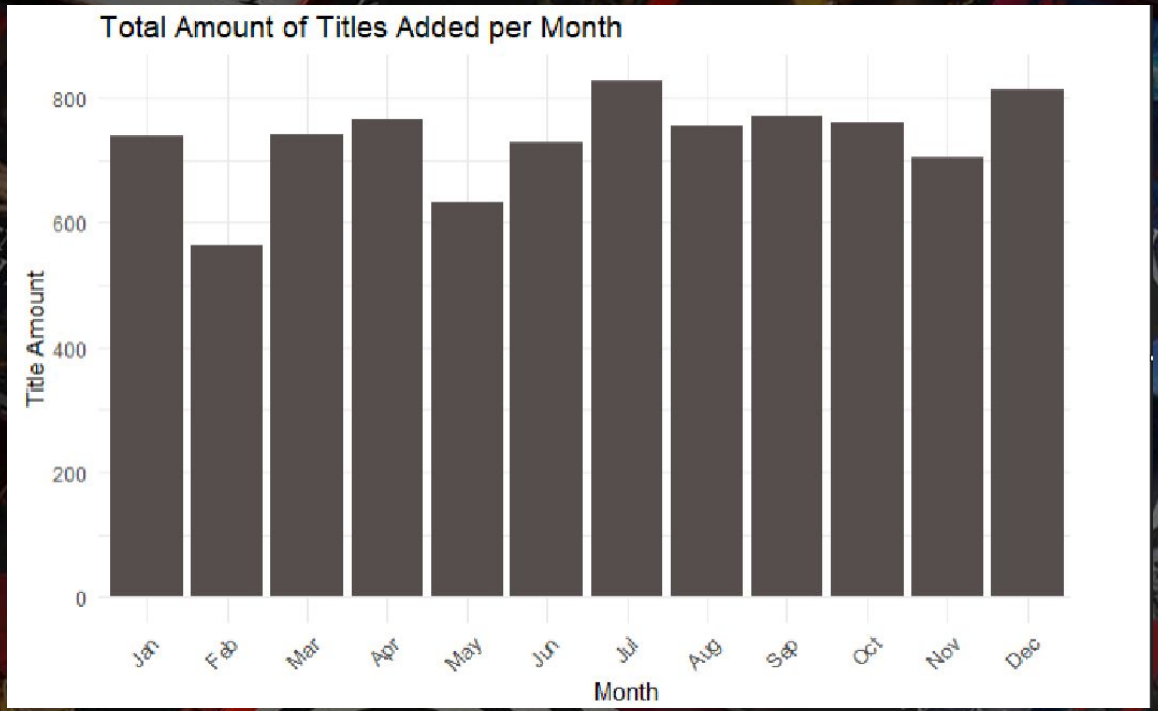
- Useful for finding positive/negative feelings in text
- Pairs words with a dictionary to give count of negative and positive words
- Assigns score to each comment from [-1 , 1]

Example(Jaws)



WHEN

Total Amount Projects Added in a year



Question 10: How does the time of year influence movie/show preferences?

- Overall, the distribution is normal for the most part
- July and December have the most amount of titles added
- Holidays/Vacation Seasons lead to more free time

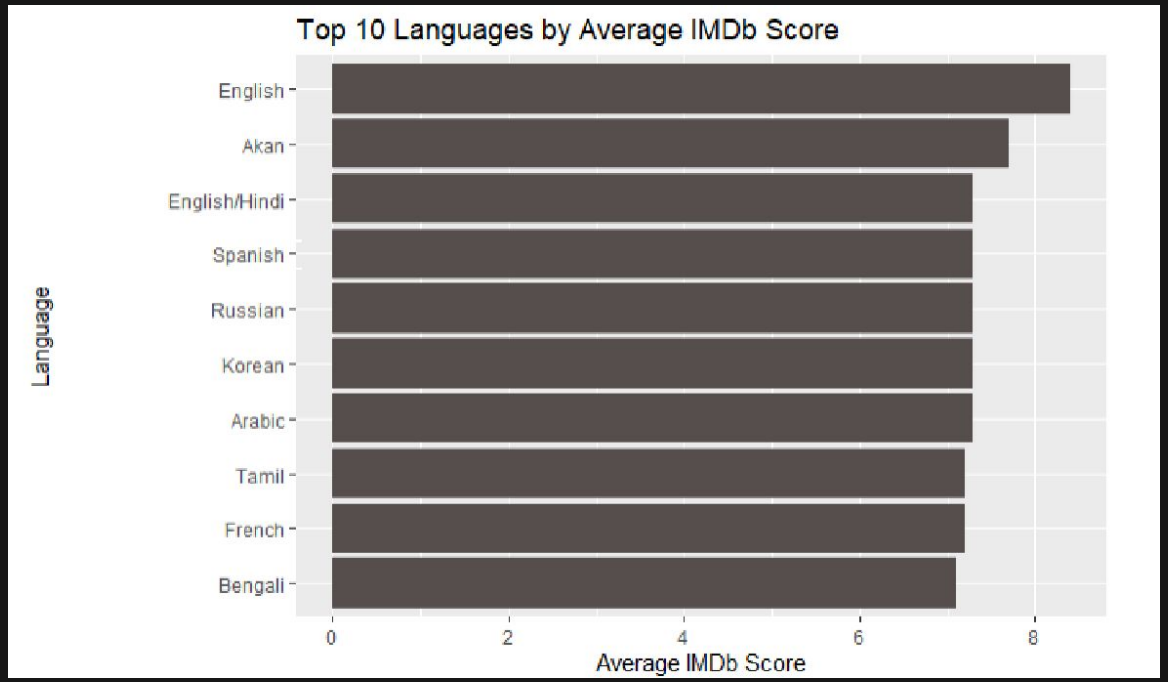
- The time of month and year leads to specific content preferences in terms of genre
- Genre preferences can relate to the time of year (July, October, December)
- The titles released in the Genre is mostly common for each month

Total Amount Projects Added in a year

MONTH	GENRE	TITLES PER GENRE
January	Comedy	4
February	Political Thriller	6
March	Romance	7
April	Science Fiction	5
May	Action	8
June	Crime	7
July	Action	5
August	Drama	6
September	Comedy	7
October	Crime	6
November	Drama	6
December	Christmas Action	6

WHERE

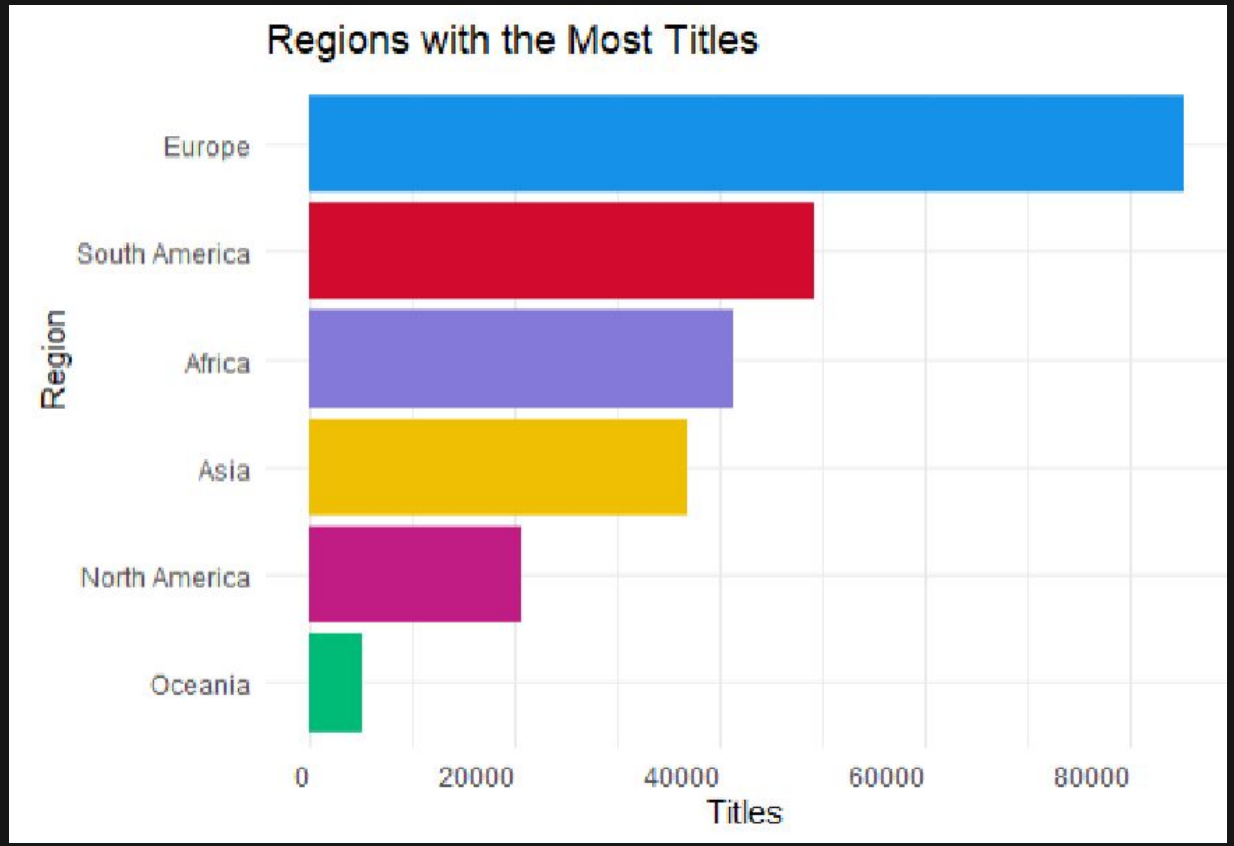
Top Primary Languages by Average IMDb Score



Question 11: Is there a preference among customers for titles produced in specific regions?

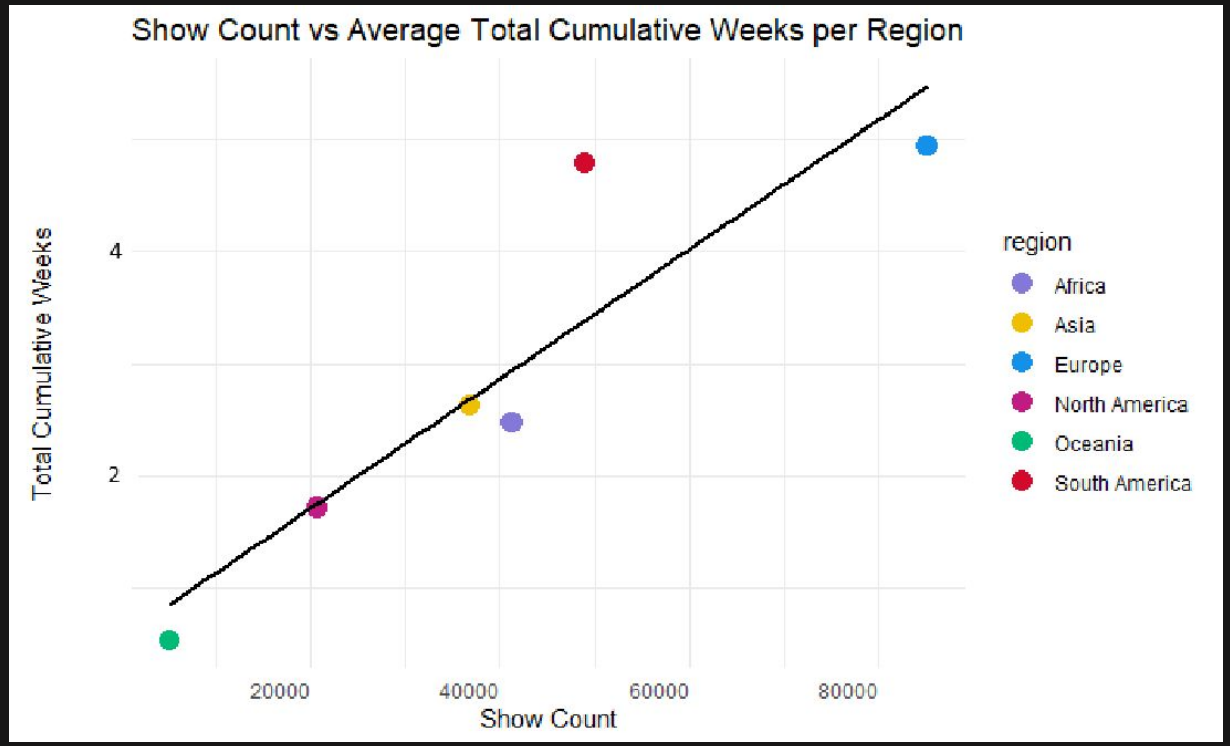
- Language accessibility and cultural familiarity affect customer preferences
- English and Akan titles have the most average ratings
- French and Bengali are among the lowest

Regions with the Most Netflix Titles made



- Europe dominates with over 80,000 titles produced
- Oceania has the least amount of titles produced with about 5000
- People want to watch more content
- More content being produced can be result of demand

Netflix Shows Count vs Average Cumulitve Weeks per Region



- Each region contains the average cumulative weeks in the top 10 of all projects produced
- Europe has the most average cumulative weeks based on teir number of shows, but not far behind from South America
- There is a strong correlation
- More content produced -> More cumulative weeks the content would show in the top 10

RESOURCES

Content Details Dataset:
https://www.kaggle.com/datasets/shivamb/netflix-shows/data

Cast Information Dataset: https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

Content Ratings and Genre Dataset:
https://www.kaggle.com/datasets/thedevastator/netflix-imdb-scores

Top 10 Chart Count: https://www.kaggle.com/datasets/mikitkanakia/netflix-top-10-weekly-dataset

Global Content Dataset: https://www.kaggle.com/datasets/sujaykapadnis/official-netflix-streaming-data/data

IMDB Review Data:
https://www.imdb.com/title/tt0073195/reviews/?ref_=tt_urv_sm&sort=submission_date%2Cdesc

Text Mining assumptions:
https://www.ibm.com/topics/bag-of-words#:~:text=Bag%20of%20words%20assumes%20words%20are%20independent%20of,not%20account%20for%20correlations%20in%20usage%20between%20words