Analysis within the film Industry

By: Kiran Patel, Margaret Wade, Alka Kumari and Nyi Minn Thant (group 4)

Introduction

The film industry has consistently commanded the spotlight within the entertainment industry. A report from the Motion Picture Association revealed a global box office revenue of $42.2 billion in 2019, a testament not only to its widespread popularity but also to its profound influence on the global economy (Escandon, 2020). Navigating this dynamic industry is a brave task, offering rewards for all stakeholders, including both viewers and filmmakers. Given such a large industry, there are countless facets that one could look into but perhaps one of the most important questions to look at is what goes into a successful film and how do these successful films represent society? We set out to study and understand what factors make a film successful and if the success of a movie can be predicted using the numerous factors surrounding movie production, as well as how the evolution of the movie industry over time reflects and impacts cultural changes and societal norms.

While we initially aimed to design a film recommendation model if possible, after looking into the article called “A Recommendation for Predicting Movie Ratings Using A Big Data Approach” developed by Awan et al. (2021), we felt the extent of mathematical knowledge and time commitment required to build such a model goes beyond the scope of the course objective. We then decided to take the project to a more analytical direction by using number-based and sentiment-based approaches to answer five different questions designed to gain specific insight into various aspects of the film industry. The scope of these questions include film success predictions, exploring genre trends through time and languages, and assessing whether the film industry has changed to reflect societal norms from gender and language perspectives. These questions are seen below:

Numbers Based:

1. What variables help predict movie ratings and popularity?
2. How do genre trends change over time with respect to rating and popularity scores?
3. What are the most popular genres based on the language of the film? (Global approach)

Sentiment Based:

4. How do the most popular words per genre change over time when looking specifically at tagline? Does this mirror societal change?
5. How does casting impact ratings and popularity? Has casting changed over time to represent a more inclusive society?

Data

For this analysis, we used two different datasets. The first dataset we used is from Kaggle called “The Movies Dataset.” This dataset contained a primary file, ‘movies_metadata.csv,’ which provided metadata for over 45,000 movies ranging from 1874 to 2020. This file included metrics such as release date, movie name, movie ID, budget, revenue, languages, production companies and countries. Alongside this dataset was another file called ‘credits.csv,’ which contained JSON formatted cast and crew data, each storing information about individuals such as their name and gender, alongside movie ID to identify the associated movie. There was another file called “ratings.csv” which captured important metrics when it came to the way the movie was scored/rated from the audience. Lastly, there was a file called “links.csv” which helps link together all of these files to create a comprehensive dataset. While Kaggle calls this dataset “The Movies Dataset” as a whole, there are 7 smaller files that make up the total dataset. In this specific project, we use four of these files.

The second dataset that we used was used for the sentiment analysis portion of the project. The dataset was called “Movie Script Corpus” and contained 3,000 different movie scripts. This file contained the same identifier included in the script file names, alongside information such as release year, user rating, and producers.

Two main values that will be discussed in this paper are Popularity and Rating gained from the Movies Dataset from Kaggle. They are defined here below:

1. Vote Rating: From 0 to 5 with 0.5 increments where there are multiple ratings by different users per movie and average rating is calculated
2. Popularity score: Calculated by IMDB based on votes, views, times marked as favorite, and number of times added to watchlist.

Part 1 - Numbers Based Approach

Question 1 – Methodology

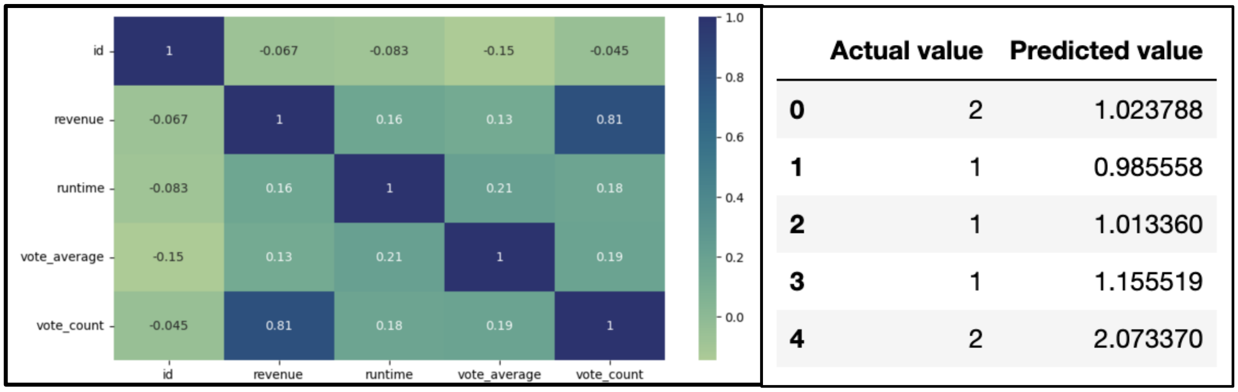
Our first question aimed to look at what factors could influence popularity and ratings. We decided to do this by running many of the data analysis tests that we learned this semester such as multiple linear regression, predicted values, correlation, and coefficient of determination. The first step was to import and merge the data. As mentioned in the data section, there were multiple files nested within the larger dataset. Within this question, the movies_metadata.csv file and the ratings.csv file were needed. The first step was merging these two files on the Movie ID column. This process was done twice resulting in two different data sets—one that had the data for the popularity scores and the second that contained the rating scores. Once the data was merged, we needed to clean the data to prepare it for analysis. For the popularity data, there were 371 movies removed and for the ratings data, over 5,000 movies were removed which included duplicate and null values.

After merging and cleaning the dataset, a three-fold analysis was conducted. Multiple linear regression analysis was performed separately for popularity and ratings, utilizing budget, revenue, run time, vote average, and vote count as independent variables. Next, predicted values were generated to see how accurate the model from the Multiple Linear Regression truly is. Lastly, an R-squared value was calculated to gauge how well popularity and rating were impacted by different variables. Finally, a correlation analysis was executed to assess the strength and direction of relationships between variables surrounding rating and popularity. These analyses collectively offer a comprehensive insight into the dataset setting up the basis of this analysis.

Question 1 - Evaluation/Results

For the popularity model, the equation **Popularity = 1.37e-8 Budget + 1.19e-8 Revenue + 1.06e-2 Run Time + 0.43 Vote Average + 4.51e-3 Vote Count + 0.95** presents interesting insights. Despite the small coefficients, the profound influence of budget and revenue—often in the multimillion or billion-dollar range—becomes apparent, challenging initial perceptions. This model achieved an R-squared value of 77.39%, which provides some insight into how popularity is impacted by these five different variables.

Figure 1. The correlation and predicted values of the data for Popularity where 0 = Budget, 1 = Revenue, 2 = Run Time, 3 = Vote Average, 4 = Vote Count

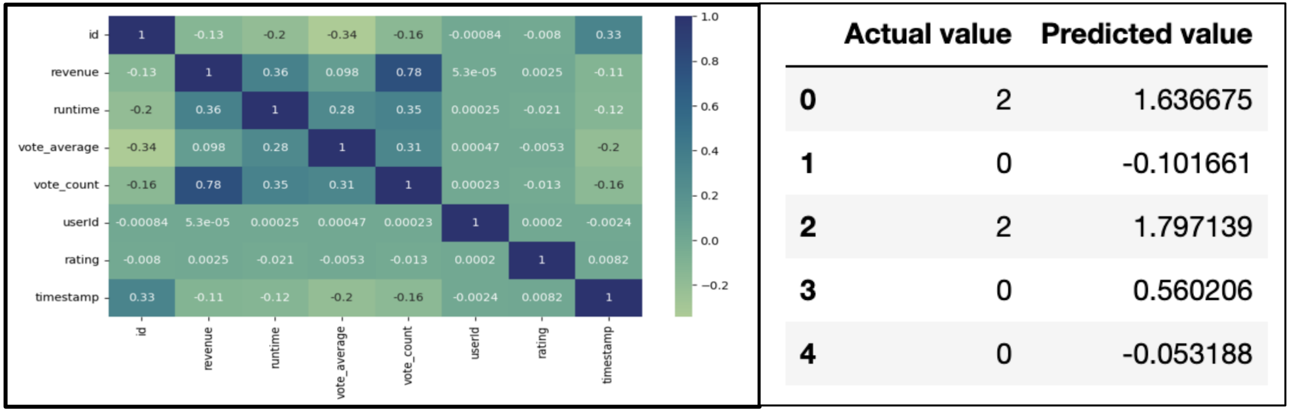


	id	revenue	runtime	vote_average	vote_count
id	1	-0.067	-0.083	-0.15	-0.045
revenue	-0.067	1	0.16	0.13	0.81
runtime	-0.083	0.16	1	0.21	0.18
vote_average	-0.15	0.13	0.21	1	0.19
vote_count	-0.045	0.81	0.18	0.19	1

	Actual value	Predicted value
0	2	1.023788
1	1	0.985558
2	1	1.013360
3	1	1.155519
4	2	2.073370

The second model surrounding ratings gave the following output. **Rating = -1.62e-3 Vote Avg + -9.55e-6 Vote Count + 1.74**. Although the predicted values for vote average and vote count deviated a bit from the actual values, the overall R-squared value was impressive at 93.22%.

Figure 2. The correlation and predicted values of the data for Rating where 0 = Budget, 1 = Revenue, 2 = Run Time, 3 = Vote Average, 4 = Vote Count



	id	revenue	runtime	vote_average	vote_count	userId	rating	timestamp
id	1	-0.13	-0.2	-0.34	-0.16	-0.00084	-0.008	0.33
revenue	-0.13	1	0.36	0.098	0.78	5.3e-05	0.0025	-0.11
runtime	-0.2	0.36	1	0.28	0.35	0.00025	-0.021	-0.12
vote_average	-0.34	0.098	0.28	1	0.31	0.00047	-0.0053	-0.2
vote_count	-0.16	0.78	0.35	0.31	1	0.00023	-0.013	-0.16
userId	-0.00084	5.3e-05	0.00025	0.00047	0.00023	1	0.0002	-0.0024
rating	-0.008	0.0025	-0.021	-0.0053	-0.013	0.0002	1	0.0082
timestamp	0.33	-0.11	-0.12	-0.2	-0.16	-0.0024	0.0082	1

	Actual value	Predicted value
0	2	1.636675
1	0	-0.101661
2	2	1.797139
3	0	0.560206
4	0	-0.053188

While this initial analysis did have some interesting takeaways, the findings were not extremely transformative or important enough to continue exploring. After diving into this question, we recognized the need for a more comprehensive look into the industry, which is what the rest of the paper is designed to do. For the remainder of the paper, we turn our focus to the influential realms of popular genres and casting choices and how this can impact popularity and shift away from a the numerical datapoints. Our aim is to uncover different elements that significantly impact both popularity scores and ratings making this study more comprehensive and holistic as we approach it from multiple different lenses.

Question 2 - Methodology

Embarking into the exploration of movie genres, this question aims to answer how trends surrounding movie genres has changed over time. Films and movies have become such a hallmark of today's cultural identity, and this question is designed to provide a glimpse into the cultural past to investigate if and how cultural changes of the past are reflected in the film art form. This will help identify if there are any genres that are timeless and/or if there are any genres that have emerged with more success in recent times. The popularity score and average rating metrics will be used to evaluate the success of movies within a certain time period. Since each movie can have more than one genre, the occurrences of each genre in these top 50 movies for every decade will be counted and ranked. We will showcase the top 5 genres most frequently found among the top 50 movies for each decade, based on each evaluation metric, in order to address this inquiry comprehensively.

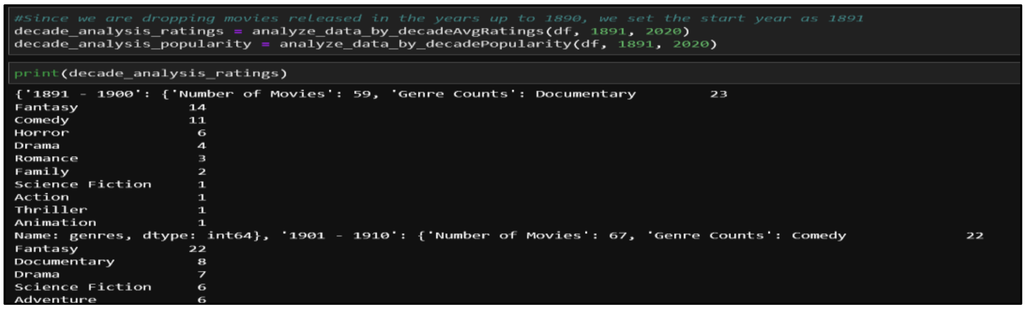
The data files used are "movies_metadata.csv" for the release date, popularity score and genres, "ratings.csv" for rating scores, and "links.csv" that provides id translation key since "movies_metadata.csv" and "ratings.csv" uses different type of id numbers for the same movies. Links data has three columns: movieId(also used by the ratings dataset), imdbId(used by metadata), and tmdbId(used by metadata as well but named as "id" there). These three columns were tested for presence of duplicate values and null values. tmdbId column has both empty rows and duplicates while both movieId and imdbId columns do not have any empty or duplicate values. For this reason, the ratings dataset's movieId column will be translated to imdbId by merging with the links dataset. Then the new ratings dataset will be merged with the metadata dataset where the rows have the same imdbId.
Before merging the datasets, the "movies_metadata.csv" is loaded as a pandas dataframe (referred to as movies_metadata from now on) and cleaned. Rows having duplicate values on the imdb_id column are identified, and extra rows were removed along with rows having null values in this column.

The "ratings.csv" dataset has rating scores given by multiple users to each movie in the data set. The average rating score is calculated by grouping the rows by unique movieId and calculating the average. Then, the imdbId from the links dataset for the respective movieIds are added as an extra column by merging the two data sets. The resulting dataframe has three columns "movieId", "avg_rating" and "imdbId".

By this point, the ratings dataset and movies_metadata had different numbers of rows, suggesting that there are movies that are in one dataset but missing in the other. These are identified and dropped from both datasets. This results in ratings dataset and movies_metadata having the same length and the same movies. Then, movies_metadata is cleaned further by selecting only "imdb_id", "release_date", "genres" and "popularity" columns only. These two datasets are then merged. This merged dataframe is then cleaned further by dropping rows with empty "release_date" and "genre" values to produce the finalized clean data ready for analysis. Within this merged dataset, the release dates ranged from 1874 to 2020. To ensure even decade splits, the decision to exclude movies from before the year 1881 was made, giving us movies released in 14 decades from 1881 to 2020 to be explored. However, in another step, we found that there are only 6 movies recorded with release dates between 1881 to 1890. So, this decade was dropped as well, resulting in 13 decades worth of movies to explore.

A custom python function for parsing genres was created since the genres are stored in json format in the data set. This function was then embedded into two new custom functions (one for popularity score and one for average rating) that looped through movies by decades, selected the top 50 movies with the highest respective metric score type, parsed, exploded and counted the genre types. The functions take a dataframe, start and end years as inputs and ran for years 1891 to 2020. The output is the number of movies recorded in each decade with the frequency of genre within the top 50 movies ranked from highest genre occurrence count to lowest.

Figure 3. Sample output of the function: showing only the first decade and part of second decade for genre occurrence ranking of top 50 movies with highest rating in each decade



```
#Since we are dropping movies released in the years up to 1890, we set the start year as 1891
decade_analysis_ratings = analyze_data_by_decadeAvgRatings(df, 1891, 2020)
decade_analysis_popularity = analyze_data_by_decadePopularity(df, 1891, 2020)
```

```
print(decade_analysis_ratings)
```

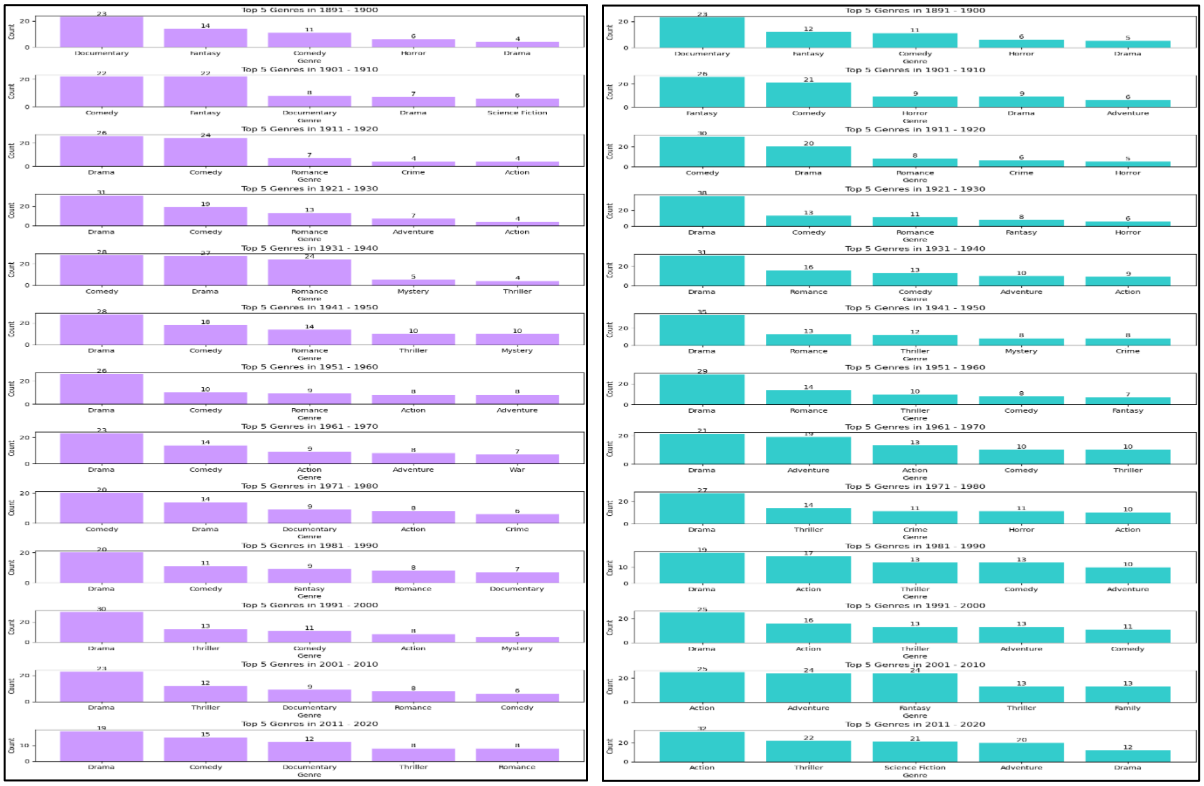
```
{'1891 - 1900': {'Number of Movies': 59, 'Genre Counts': Documentary        23
Fantasy            14
Comedy             11
Horror              6
Drama               4
Romance             3
Family              2
Science Fiction     1
Action              1
Thriller            1
Animation           1
Name: genres, dtype: int64}, '1901 - 1910': {'Number of Movies': 67, 'Genre Counts': Comedy            22
Fantasy            22
Documentary         8
Drama               7
Science Fiction     6
Adventure           6
```

<u>Question 2 - Evaluation/Results</u>

We are able to identify a few timeless genres like Drama, Comedy and to a lesser extent Romance. These three genres are present in the top 5 of almost every decade in both rating and popularity. Due to these genres, the trends in genre success seems to stay mostly unchanged throughout the 13 decades. It was surprising to learn that science fiction movies were being made with success as early as the 1900s and people seem to take interest in Horror movies starting from the 1890s. Genres like Mystery, Crime and Thriller start to gain traction starting from the 1940s in both ratings and popularity.

Another interesting takeaway is that in movies with the highest popularity score, Action and Adventure genres begin to dominate starting from the 1980s with the Action genre taking the top spot firmly in the 2000s and 2010s. Fantasy and Science Fiction genres also made a major comeback in these two decades. This coincides with the gaining popularity of superhero movies like the Marvel Cinematic Universe movies and the massive marketing powers behind these movies in recent years. However, these trends are not reflected in the top-rated movies of the same decades. The "more serious" genres like Drama, Thriller, Romance and Comedy still dominate the top genre positions in all decades. A possible explanation for this is that while ratings measure how well a movie is made, popularity score also considers the interest generated around it. While the dataset is not explicit about how exactly the popularity score is calculated, it is mentioned the most impactful factors include how often a movie is added to a "favorites" list and/or "watchlist", how often a movie is voted for, and how often its info page is viewed. Due to the marketing power behind recent Action and superhero movies (which are considered to fall within the Action/Science Fiction/Fantasy genres), they are able to generate high popularity scores even if a portion of such movies fail to receive good ratings.

Figure 4. Visualization of five most commonly occurring genres in the highest scoring 50 movies of every decade from 1891 to 2020, based on average ratings (left/purple) and based on popularity score (right/cyan)



A limitation of this study is that the number of movies recorded in each decade are not consistent. As seen in figure 3, the number of movies in the first two decades are 59 and 67 respectively. This makes the genres more representative of all the movies made in those years rather than the top few selections. However, the movie count exploded in the following decades up to 10 thousand. Selecting a top percentage of the total number of movies in the decade might help normalize the results but that method still would not work for decades with really low numbers like the first two decades. The extreme values will still affect the results in undesired ways.

Another limitation is that metrics like “watchlist” adding, popularity scores and user ratings are very modern concepts. People in the past before the conception of these measurements had no way of participating in having their opinions recorded this way. Because of this, rating and popularity scores of older movies are diluted by opinions of people who lived and seen the movies in more recent times. This might mean that the results of this analysis aren’t very representative of the opinions expressed at the time of older movies’ release. One way to work around this is to identify metrics that measure the success of a movie and also minimizes the dilution of opinions. One such possible metric is revenue since movies make most of the revenue around the time of release (Arkenberg et al. 2020). Even then, it is impossible to completely isolate the impact of future earnings, only minimized. This question can be expanded by identifying additional metrics and then carrying out the analysis on most successful movies based on those new identifiable metrics.

Question 3 - Methodology

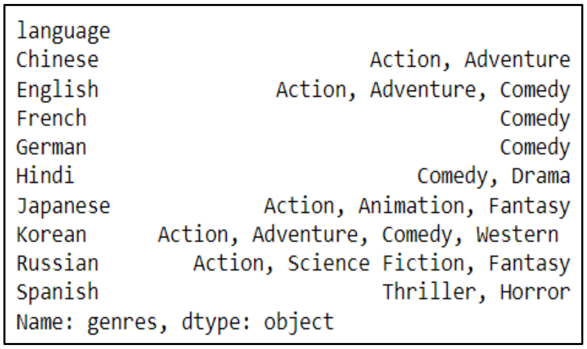
While the previous question provided context into popular genres over time, it did not dive into how this could differ geographically. Genres such as Thriller and Drama may enjoy popularity confined to specific regions and linguistic communities but not in others. For example, when one thinks of Bollywood, they likely are not grouping those films under a thriller/horror category because that is not something they are known for. Instead, Bollywood is renowned for its iconic romance narratives filled with singing and dancing. While question two looks at the top genres over time, it is equally as important to identify how cinematic preferences may differ across the world.

By looking into popular genres worldwide, directors can gain insights into cultural nuances and regional variations in film preferences. Filmmakers can utilize this information as a guide when creating new content. For instance, filmmakers may decide to produce more content in a particular genre if it regularly does well in that language in order to optimize revenue. Similarly, investors can use this information to make informed investment decisions. To address the 'popular genres based on language', an exploratory data analysis was performed to understand the structure and characteristics of the movie dataset. The intent was to understand the genres, languages, and revenue from the dataset. The files were thoroughly cleaned to remove the duplicate and null values from the dataset.

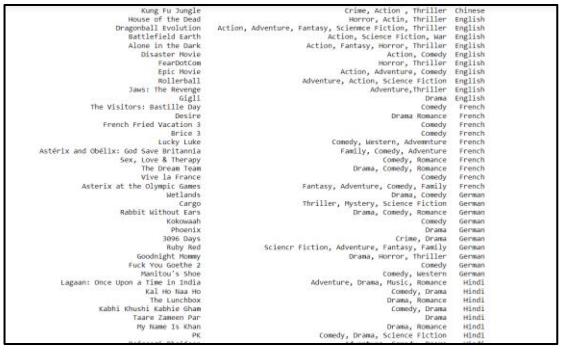
Question 3- Evaluation/Results

Revenue analysis: The first step was to analyze the top 10 languages based on movie revenue. Furthermore, for each language, the top 10 movies were selected as per the revenue.

Figure 5. The top 10 languages as per revenue / Figure 6. The top 10 movies in each language as per revenue



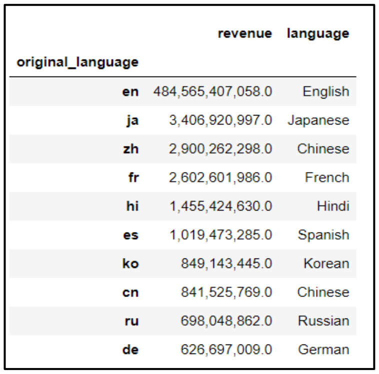
```
language
Chinese                          Action, Adventure
English                  Action, Adventure, Comedy
French                                      Comedy
German                                      Comedy
Hindi                                Comedy, Drama
Japanese                  Action, Animation, Fantasy
Korean         Action, Adventure, Comedy, Western
Russian        Action, Science Fiction, Fantasy
Spanish                            Thriller, Horror
Name: genres, dtype: object
```



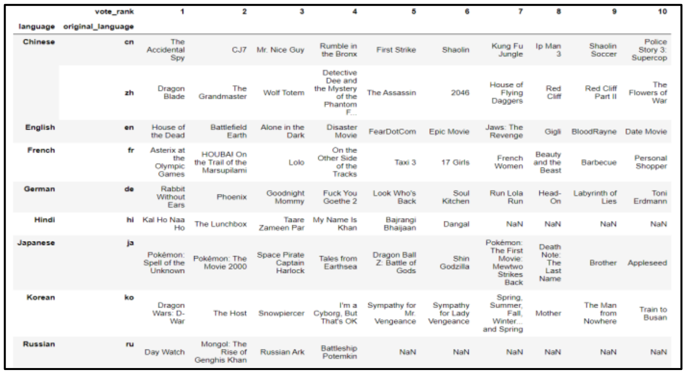
```
                        Kung Fu Jungle                          Crime, Action , Thriller  Chinese
                    House of the Dead                            Horror, Actin, Thriller  English
                  Dragonball Evolution  Action, Adventure, Fantasy, Sciemce Fiction, Thriller  English
                  Battlefield Earth                         Action, Science Fiction, War  English
                   Alone in the Dark                    Action, Fantasy, Horror, Thriller  English
                      Disaster Movie                                     Action, Comedy  English
                          FearDotCom                                   Horror, Thriller  English
                          Epic Movie                          Action, Adventure, Comedy  English
                          Rollerball                  Adventure, Action, Science Fiction  English
                    Jaws: The Revenge                              Adventure,Thriller  English
                               Gigli                                            Drama  English
           The Visitors: Bastille Day                                          Comedy   French
                              Desire                                   Drama Romance   French
            French Fried Vacation 3                                           Comedy   French
                             Brice 3                                           Comedy   French
                          Lucky Luke                     Comedy, Western, Advemnture   French
Astérix and Obélix: God Save Britannia                     Family, Comedy, Adventure   French
                  Sex, Love & Therapy                               Comedy, Romance   French
                       The Dream Team                         Drama, Comedy, Romance   French
                       Vive la France                                       Comedy   French
           Asterix at the Olympic Games          Fantasy, Adventure, Comedy, Family   French
                            Wetlands                                 Drama, Comedy   German
                               Cargo            Thriller, Mystery, Science Fiction   German
                   Rabbit Without Ears                      Drama, Comedy, Romance   German
                            Kokowaah                                        Comedy   German
                             Phoenix                                         Drama   German
                           3096 Days                                  Crime, Drama   German
                            Ruby Red    Sciencr Fiction, Adventure, Fantasy, Family   German
                      Goodnight Mommy                      Drama, Horror, Thriller   German
                     Fuck You Goethe 2                                      Comedy   German
                       Manitou's Shoe                              Comedy, Western   German
     Lagaan: Once Upon a Time in India          Adventure, Drama, Music, Romance    Hindi
                        Kal Ho Naa Ho                                Comedy, Drama    Hindi
                        The Lunchbox                               Drama, Romance    Hindi
               Kabhi Khushi Kabhie Gham                              Comedy, Drama    Hindi
                     Taare Zameen Par                                       Drama    Hindi
                      My Name Is Khan                              Drama, Romance    Hindi
                                   PK               Comedy, Drama, Science Fiction    Hindi
```

The 'genres' field for each language's top 10 films were also examined to determine the most popular genres in the specific language. This is seen below.

Figure 7. Most common genres that a movie falls in / Figure 8. Most popular genres used by the film industry for the top 10 languages



	revenue	language
original_language		
en	484,565,407,058.0	English
ja	3,406,920,997.0	Japanese
zh	2,900,262,298.0	Chinese
fr	2,602,601,986.0	French
hi	1,455,424,630.0	Hindi
es	1,019,473,285.0	Spanish
ko	849,143,445.0	Korean
cn	841,525,769.0	Chinese
ru	698,048,862.0	Russian
de	626,697,009.0	German



	vote_rank	1	2	3	4	5	6	7	8	9	10
language	**original_language**										
Chinese	**cn**	The Accidental Spy	CJ7	Mr. Nice Guy	Rumble in the Bronx	First Strike	Shaolin	Kung Fu Jungle	Ip Man 3	Shaolin Soccer	Police Story 3: Supercop
	zh	Dragon Blade	The Grandmaster	Wolf Totem	Detective Dee and the Mystery of the Phantom F...	The Assassin	2046	House of Flying Daggers	Red Cliff	Red Cliff Part II	The Flowers of War
English	**en**	House of the Dead	Battlefield Earth	Alone in the Dark	Disaster Movie	FearDotCom	Epic Movie	Jaws: The Revenge	Gigli	BloodRayne	Date Movie
French	**fr**	Asterix at the Olympic Games	HOUBA! On the Trail of the Marsupilami	Lolo	On the Other Side of the Tracks	Taxi 3	17 Girls	French Women	Beauty and the Beast	Barbecue	Personal Shopper
German	**de**	Rabbit Without Ears	Phoenix	Goodnight Mommy	Fuck You Goethe 2	Look Who's Back	Soul Kitchen	Run Lola Run	Head-On	Labyrinth of Lies	Toni Erdmann
Hindi	**hi**	Kal Ho Naa Ho	The Lunchbox	Taare Zameen Par	My Name Is Khan	Bajrangi Bhaijaan	Dangal	NaN	NaN	NaN	NaN
Japanese	**ja**	Pokémon: Spell of the Unknown	Pokémon: The Movie 2000	Space Pirate Captain Harlock	Tales from Earthsea	Dragon Ball Z: Battle of Gods	Shin Godzilla	Pokémon: The First Movie: Mewtwo Strikes Back	Death Note: The Last Name	Brother	Appleseed
Korean	**ko**	Dragon Wars: D-War	The Host	Snowpiercer	I'm a Cyborg, But That's OK	Sympathy for Mr. Vengeance	Sympathy for Lady Vengeance	Spring, Summer, Fall, Winter... and Spring	Mother	The Man from Nowhere	Train to Busan
Russian	**ru**	Day Watch	Mongol: The Rise of Genghis Khan	Russian Ark	Battleship Potemkin	NaN	NaN	NaN	NaN	NaN	NaN

Examining the language patterns within different film genres provides valuable insights for producers aiming to connect with a diverse global audience. If a director wants to reach a widespread audience, perhaps Action and Adventure is the way to go as it seems to have a universal appeal. Comedy, for the most part, has the same effect but it leans more towards English, French, German and Hindi markets. Other areas of the world are very niche like Russian audiences showing a preference for Action and Science Fiction Fantasy and Spanish viewers having a particular appeal for Thriller/Horror films. Other audiences have much wider preferences like the Japanese audiences who demonstrate a varied interest in Action, Animation, and Fantasy genres, signaling a potential openness to creative and animated storytelling. Similarly, Korean films display diverse genre preferences, encompassing Action, Adventure, and Comedy, hinting at the potential success of genre-blending in this market.

In conclusion, it is important for producers to understand the importance and impact of cultural norms and preferences as it can help them strategically craft films for specific audiences.

Part 2: Sentiment based Approach

The first part of this project looked into genres and their popularity in depth. The second portion of this paper will delve into more of a sentiment-based analysis when it comes to the film industry looking into specific words used within genres while also taking a look at casting over time and throughout different sectors of the industry. The purpose here is to better understand if the film industry has done a sufficient job at reflecting the cultural desires/the cultural norms of the time.

Question 4 - Methodology

The first part of the sentiment analysis portion was to conduct a comprehensive analysis of movie taglines to uncover associations between genres and the frequency of specific words. We chose tagline here instead of looking at the entire script/synopsis as we wanted to approach this from a marketing perspective, again trying to stand in the shoes of a director/producer. A tagline is one of the very first consumer touchpoints in terms of a new movie release which results in a wide variety of expectations and perceptions about a film. Because of this, we deemed this a very important measure in terms of genre analysis as a tagline is simple but effective in telling the theme of a film.

We initiated this analysis by loading it into the Kaggle movie dataset focusing only on the "tagline" and "genres" columns. We then conducted a genre-specific exploration of movie taglines, extracting and processing words to identify the top 20 frequently occurring words for each genre. This was stored in a dataframe which organized the information with columns denoting the movie genres, top 20 words and the frequency that the word appeared. Frequency was kept to better understand the amount of times these words were said and help differentiate the top few words from the 19th and 20th more popular words. This approach allowed for a comprehensive overview of the characteristics associated with different movie genres.

To share our insights, we exported the final DataFrame to a CSV file, named 'Genre_Word_Frequencies.csv.' We put the resulting export into Tableau where we were able to visualize these top words per genre to allow for viewers to visualize the meaning of this output rather than simply digesting the python result. We repeated this process two different times. The first was to get most utilized words by genre and decade and the second was to get most utilized words by popularity score (grouping by every 10).

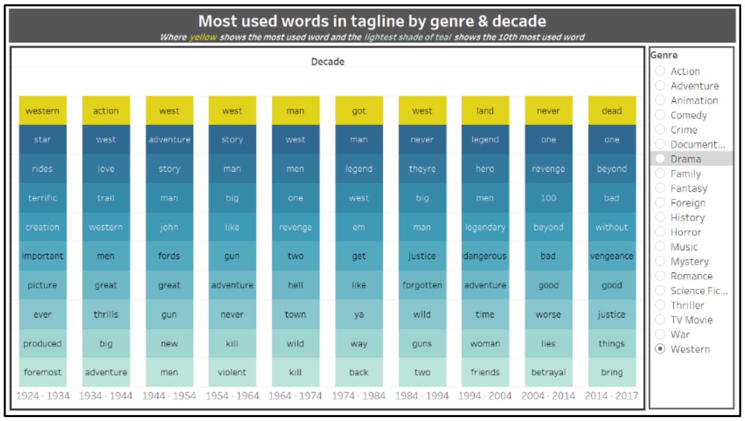
<u>Question 4- Evaluation/Results</u>

Case 1: By Decade and Genre

For this section in particular, we wanted to take a look at three different genres: Western, Romance and Drama. Western was popular in the early 1900s while Romance has remained in the mix throughout. Drama is one of the timeless genres that we were able to find from the results of question 2 so we wanted to see if the themes within Drama have remained consistent or have changed throughout its reign as the most popular. Using the help of AI technology, we were able to get the connotations of the most commonly used words to help create a thematic analysis.

a) Within the Western genre, the 1924-1934 theme centers on the American West, combining expansive landscapes, human presence, and a blend of civilization and wilderness. The mention of "guns" suggests ruggedness, while "forgotten" implies neglect, and "two" hints at duality. The 2014-2017 theme explores morality, justice, and consequences, with a focus on negativity, moral righteousness, and a transcendent aspect beyond the ordinary. From this analysis, we can see the Western genre has deviated a bit from its historical roots which can explain why it is not as popular as it once used to be.

Figure 8: Most used words in tagline when Genre is Western



Most used words in tagline by genre & decade

Where yellow shows the most used word and the lightest shade of teal shows the 10th most used word

Decade

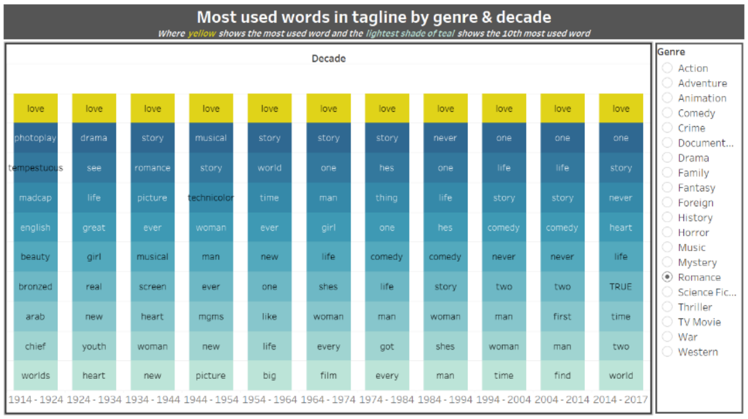
western	action	west	west	man	got	west	land	never	dead
star	west	adventure	story	west	man	never	legend	one	one
rides	love	story	man	men	legend	theyre	hero	revenge	beyond
terrific	trail	man	big	one	west	big	men	100	bad
creation	western	john	like	revenge	em	man	legendary	beyond	without
important	men	fords	gun	two	get	justice	dangerous	bad	vengeance
picture	great	great	adventure	hell	like	forgotten	adventure	good	good
ever	thrills	gun	never	town	ya	wild	time	worse	justice
produced	big	new	kill	wild	way	guns	woman	lies	things
foremost	adventure	men	violent	kill	back	two	friends	betrayal	bring
1924 - 1934	1934 - 1944	1944 - 1954	1954 - 1964	1964 - 1974	1974 - 1984	1984 - 1994	1994 - 2004	2004 - 2014	2014 - 2017

Genre

○ Action
○ Adventure
○ Animation
○ Comedy
○ Crime
○ Document...
○ Drama
○ Family
○ Fantasy
○ Foreign
○ History
○ Horror
○ Music
○ Mystery
○ Romance
○ Science Fic...
○ Thriller
○ TV Movie
○ War
◉ Western

b) Within the Romance genre, 1914-1924 theme revolves around what seems to be a foreign plot. It mentions words like “worlds”, “Arab”, “bronzed” which references things from around the world. Throwing in words like English shows that it could potentially be a plot about lovers from across the world. 2014-2017 theme also centers on love and relationships, emphasizing uniqueness with "one" and authenticity with "true." "Story" suggests a narrative or journey, "heart" adds an emotional dimension, and "world" expands the context. "Two" implies duality or partnership. The first set hints at a foreign concept with elements of love while the second set leans toward a narrative focusing on love, family and the search for meaning in life.

Figure 9: Most used words in tagline when Genre is Romance



Most used words in tagline by genre & decade

Where yellow shows the most used word and the lightest shade of teal shows the 10th most used word

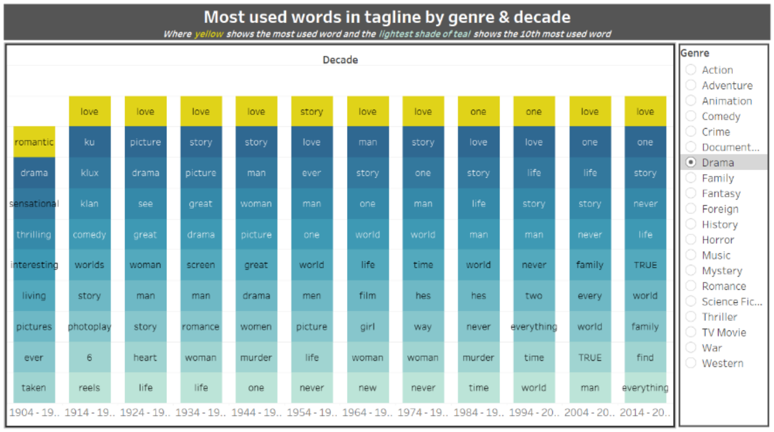
Decade

1914 - 1924	1924 - 1934	1934 - 1944	1944 - 1954	1954 - 1964	1964 - 1974	1974 - 1984	1984 - 1994	1994 - 2004	2004 - 2014	2014 - 2017
love	love	love	love	love	love	love	love	love	love	love
photoplay	drama	story	musical	story	story	story	never	one	one	one
tempestuous	see	romance	story	world	one	hes	one	life	life	story
madcap	life	picture	technicolor	time	man	thing	life	story	story	never
english	great	ever	woman	ever	girl	one	hes	comedy	comedy	heart
beauty	girl	musical	man	new	life	comedy	comedy	never	never	life
bronzed	real	screen	ever	one	shes	life	story	two	two	TRUE
arab	new	heart	mgms	like	woman	man	woman	man	first	time
chief	youth	woman	new	life	every	got	shes	woman	man	two
worlds	heart	new	picture	big	film	every	man	time	find	world

Genre
○ Action
○ Adventure
○ Animation
○ Comedy
○ Crime
○ Document...
○ Drama
○ Family
○ Fantasy
○ Foreign
○ History
○ Horror
○ Music
○ Mystery
● Romance
○ Science Fic...
○ Thriller
○ TV Movie
○ War
○ Western

c) Lastly, within the Drama genre, the 1904-1914 theme suggests a mystery or crime narrative with elements of enduring love and personal relationships. It centers around a singular focus, potentially involving a dark plot with murder. In contrast, the 2014-2017 theme revolves around the intertwining themes of love, family, and the search for authentic meaning in life. It emphasizes uniqueness, authenticity, and completeness, with a narrative structure that delves into familial relationships and the significance of personal quests.

Figure 10: Most used words in tagline when Genre is Drama



Most used words in tagline by genre & decade

Where yellow shows the most used word and the lightest shade of teal shows the 10th most used word

Decade

1904 - 19..	1914 - 19..	1924 - 19..	1934 - 19..	1944 - 19..	1954 - 19..	1964 - 19..	1974 - 19..	1984 - 19..	1994 - 20..	2004 - 20..	2014 - 20..
	love	love	love	love	story	love	love	one	one	love	love
romantic	ku	picture	story	story	love	man	story	love	love	one	one
drama	klux	drama	picture	man	ever	story	one	story	life	life	story
sensational	klan	see	great	woman	man	one	man	life	story	story	never
thrilling	comedy	great	drama	picture	one	world	world	man	man	never	life
interesting	worlds	woman	screen	great	world	life	time	world	never	family	TRUE
living	story	man	man	drama	men	film	hes	hes	two	every	world
pictures	photoplay	story	romance	women	picture	girl	way	never	everything	world	family
ever	6	heart	woman	murder	life	woman	woman	murder	time	TRUE	find
taken	reels	life	life	one	never	new	never	time	world	man	everything

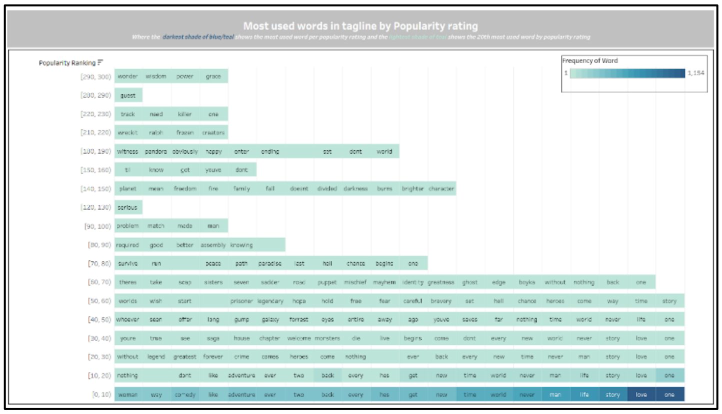
Genre
○ Action
○ Adventure
○ Animation
○ Comedy
○ Crime
○ Document...
● Drama
○ Family
○ Fantasy
○ Foreign
○ History
○ Horror
○ Music
○ Mystery
○ Romance
○ Science Fic...
○ Thriller
○ TV Movie
○ War
○ Western

To conclude, the Western genre, rooted in the American frontier, has experienced a notable shift from its historical themes, possibly contributing to a decline in popularity. On the contrary, Romance and Drama showcase a remarkable constancy in their core themes, transcending temporal boundaries. While the Western genre's deviation might explain a lessening appeal, Romance and Drama's enduring popularity lies in their ability to adapt without compromising their fundamental values. As cinematic landscapes change, these genres serve as reliable conduits for exploring universal human experiences. This analysis underscores the dynamic nature of storytelling, where certain genres stand the test of time by staying true to the core themes that resonate across generations. Ultimately, the cinematic journey through these genres becomes a reflection of societal shifts, providing a lens through which we can understand the evolving tapestry of human narratives.

Case 2: By Popularity

It is evident that movies with a rating surpassing 10 are scarce, resulting in a limited pool of options and consequently fewer associated words. One aspect that was interesting is that there are no discernible presence of words such as "woman," "women," or "girl" among the top words associated with these specific popularity scores. This shows that regardless of popularity, female centered language is hardly used within the most popular taglines, speaking to the fact that there is likely no female lead in these top films. Ultimately, the absence of a clear pattern in the word associations suggests a diverse range of themes and genres within this popularity range, making it challenging to identify specific trends or commonalities.

Figure 11: Most utilized words in tagline based on popularity score



Overall, this question helps understand popular thematic principals over time and through each genre. It also helps understand the language and themes utilized in popular movies. While interesting from a movie-watching lens, this can provide immense insight to producers/directors as they can cater themes and language to what is the most popular within each genre. One limitation in this study is the fact that many words were mentioned only 2-3 times but still made it in the top frequency. Taglines in nature are meant to be unique to catch the attention of the consumer so oftentimes words are not duplicated as much as we originally thought. Additionally, this does not necessarily dive into the cultural/linguistic aspect as all of the words are in English. However, this does provide a great stepping stone for any further research.

<u>Question 5– Methodology</u>

Due to the findings in the popularity score for question 4, we wanted to look into the hiring practices within the film industry. The fact that the word woman/women/girl/miss/she/her were not seen within the top words was a bit unsettling as it is clear that many themes are not surrounding women. What we wanted to look at here is to see different gender-related patterns in hiring practices and explore the potential influence of production companies, budgets and time on fair hiring. Our approach involved gathering information on cast genders, production year, production companies, and budget from the Kaggle movies dataset. While crew data was available, we chose to focus our investigation on consumer-facing aspects of the industry and their impact.

To process the data, we utilized PySpark and SparkSQL to clean and manipulate the "movies_metadata.csv" and "credits.csv" datasets. These datasets were merged using a common "id" column, and specific details were extracted from the "production_companies" and "cast" columns within the credits dataset, both stored in JSON format. We then created additional columns to capture the total count of female and male actors in each film which was combined into a calculated ratio of male to female actors, extracting the "gender" label and value from each "cast" entry. Emphasizing the ratio value in our analysis allowed us to account for variations in the number of actors across films, providing a relative perspective on hiring practices. After merging the two files, we created smaller datasets to answer our specific hiring questions in relation to year, decade, budget, production company, popularity and vote rating and whether either has the power to encourage or influence fair hiring.

Upon analyzing hiring practices, we then decided to look into script writing with the goal of seeing if there has been a reduction in measures which point to gender bias. When trying to figure out the best way to approach this, we decided to go with a similar method of question four by looking at the frequency of certain “word types,” in order to provide a quantitative measure of the emotional tone of the text. By examining the frequency of words seen as derogatory and negative toward women, as well as the prevalence of references to men vs women, we sought to gain insights into general attitudes and emphasis with respect to gender in the industry.

The rationale behind the usage of negative words towards women was to examine whether the prevalence of derogatory language had diminished, showing more of a positive shift in sentiments towards women in cinematic narratives. A challenge arose when attempting to access a list of academically negative terms as they were either unavailable to the public or behind paywalls. Because of this, we resorted to using our own research background by conducting a comprehensive literature review surrounding negative language towards women, compiling a list of words referenced by researchers as indicative of female-centric negativity. We then eliminated terms susceptible to diverse thematic interpretations, like “cow”, which ensured that the analysis remained objective. Our final list was comprised of 65 terms.

In examining general gender references, our pronoun lists served a dual purpose. First, this provided insight into the portrayal and discourse surrounding male and female characters. Second, this list helped identify shifts in representation by assessing whether references to women or neutral descriptors began to surpass those of men. By quantifying the prevalence of words such as "man," "woman," "she," and "he," we aimed to ascertain whether scripts exhibited a balanced referencing of men and women (Harris et al.).

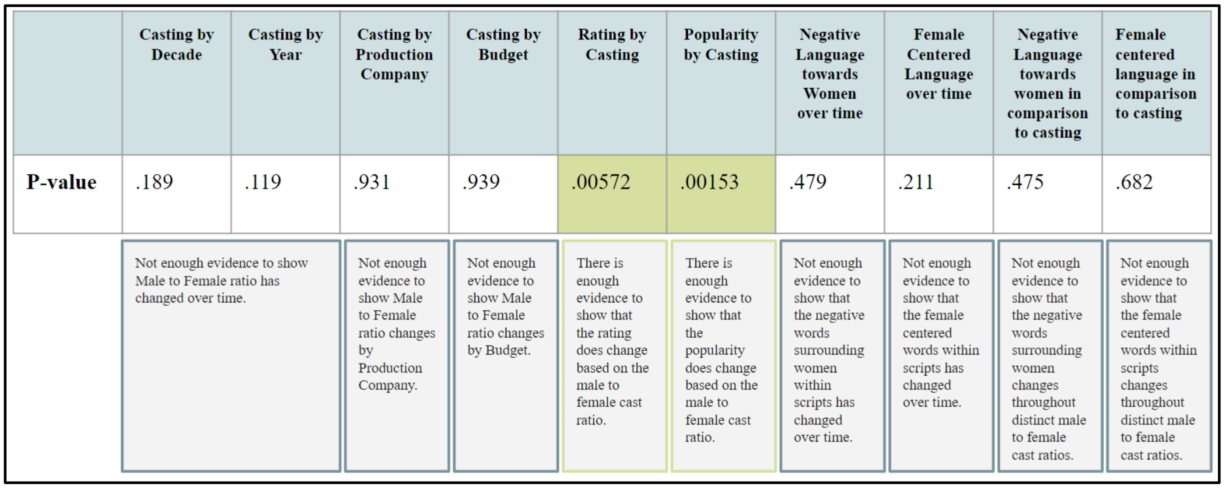
After compiling the lists, we created a python function that iterated through the scripts and tallied the occourances of female pronouns, male pronouns, neutral pronouns and negative words towards women and stored them within a dataframe. We then merged this dataframe on the Metadata ID that consisted of the male to female ratio to gain a comprehensive analysis of both script writing and casting. The resulting DataFrames, generated using PySpark, were exported as CSV files and subsequently analyzed in R. The outcomes of the analysis, presented in Figure 12, showcase the results of the ANOVAs conducted.

Question 5 - Evaluation/Results

For the most part, our analysis proved that there has not been much change for women within the film industry from both a script writing and hiring perspective. Looking then into the average male to female ratio by popularity, we found that there is a significant difference in that ratio across popularity and rating scores. Although the intricacies of this result weren't explored, the mere fact that the gender ratio does fluctuate based on popularity shows that there is a tangible impact of casting and gender on the overall reception of a movie. Even if this difference stems from unconscious bias, it is still a major factor that needs to be mitigated as films should be received equally regardless of the male to female actor count.

On the script writing side, our analysis did not reveal any significant differences in the use of language that centers women or negative language towards women when being broken down by decade and year. This further reinforces the idea that there has not been a significant change for women in the industry, both in terms of hiring and script writing. These ANVOA results are seen below.

Figure 12: P-value outputs from an ANOVA table comparing the Male to Female ratio (casting) across several different variables. The interpretation of the p-value is seen below the numbers within boxes.



	Casting by Decade	Casting by Year	Casting by Production Company	Casting by Budget	Rating by Casting	Popularity by Casting	Negative Language towards Women over time	Female Centered Language over time	Negative Language towards women in comparison to casting	Female centered language in comparison to casting
P-value	.189	.119	.931	.939	.00572	.00153	.479	.211	.475	.682
	Not enough evidence to show Male to Female ratio has changed over time.		Not enough evidence to show Male to Female ratio changes by Production Company.	Not enough evidence to show Male to Female ratio changes by Budget.	There is enough evidence to show that the rating does change based on the male to female cast ratio.	There is enough evidence to show that the popularity does change based on the male to female cast ratio.	Not enough evidence to show that the negative words surrounding women within scripts has changed over time.	Not enough evidence to show that the female centered words within scripts has changed over time.	Not enough evidence to show that the negative words surrounding women changes throughout distinct male to female cast ratios.	Not enough evidence to show that the female centered words within scripts changes throughout distinct male to female cast ratios.

Overall, this confirmed what we saw in Question 4 with the lack of female-centered terminology. This shows that the film industry is not as reflective of cultural norms and societal desires as the representation of women in the film industry has not improved or become more equitable. Further investigation could be done on how the casting changes based on genre and how that ties back into the theme.

Conclusion

In conclusion, our analysis encompassed multiple aspects of the film industry, shedding light on various factors influencing both popularity and ratings. While this study provides a beneficial foundation into many of the different variables that impact the success of a film, more research can be done to expand on what was done here. We hope this study sparks more research on genre implications, casting and how they both can impact the success of a film. Our hope is further analysis on these topics can help directors/producers streamline their script writing, hiring and marketing processes within film creation.

Acknowledgement

Due to the number of questions we wanted to answer, we split up the project by question. Margaret completed question 1, Nyi completed question 2, Alka completed question 3 and Kiran completed question 4 and 5.

Special acknowledgement to Christopher Barua and Sasha Heslin who were grouped with Kiran during a Cloud Computing course in Spring of 2023 where they dove into what is our question 5. Kiran continued to build on the solid foundation they had already built by continuing to code and come up with another way to view the data through the tagline functionality and by adding different approaches by looking at the gender ratio on popularity scores.

References:

Arkenberg, C., Cutbill, D., Loucks, J., & Westcott, K. (2020, December 10). *Digital media trends: The future of movies.* Deloitte Insights. https://www2.deloitte.com/uk/en/insights/industry/technology/future-of-the-movie-industry.html

Awan, M. J., Khan, R. A., Nobanee, H., Yasin, A., Anwar, S. M., Naseem, U., & Singh, V. P. (2021). A recommendation engine for predicting movie ratings using a big data approach. *Electronics*, *10*(10), 1215. https://doi.org/10.3390/electronics10101215

Escandon, R., (2020, March 12). *The Film Industry Made A Record-Breaking $100 Billion Last Year.* Forbes. https://www.forbes.com/sites/rosaescandon/2020/03/12/the-film-industry-made-a-record-breaking-100-billion-last-year/?sh=3a7fb76d34cd

Gaucher, Danielle, Justin Friesen, and Aaron C. Kay. "Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality." Journal of Personality and Social Psychology, vol. 101, no. 1, 2011, pp. 109-128.

Geena Davis Institute on Gender in Media. (2019). Exploring the Barriers and Opportunities for Independent Women Filmmakers. Theatrical and Streaming Audiences Research.

Harris, Chelsea A. MD, Blencowe, Natalie BMBS, PhD, and Telem, Dana A. MD, MPH. "What is in a Pronoun? Why Gender-fair Language Matters." Journal of Graduate Medical Education, vol. 7, no. 3, Sept. 2015, pp. 362-363.

Hornsby, J., & Broussard, R. (2019). A Brief Analysis of Gender Roles in Screenplays. Journal of Popular Film and Television, 47(3), 126-130.

James, Deborah. "Gender-Linked Derogatory Terms and Their Use by Women and Men." Language & Communication, vol. 14, no. 3, 1994, pp. 291–313.

Kohnen, Melanie. "Bad Language for Nasty Women (and Other Gendered Insults)." Feminist Media Studies, vol. 19, no. 5, 2019, pp. 761-764. JSTOR, www.jstor.org/stable/26746546.

Kutner, Nancy G., and Donna Brogan. "An Investigation of Sex-Related Slang Vocabulary and Sex-Role Orientation among Male and Female University Students." Journal of Marriage and the Family, vol. 41, no. 2, 1979, pp. 425-432.

Lakoff, Robin. Language and Woman's Place. New York, Harper & Row, 1975.

Lauzen, M. M. (2019). It's a Man's (Celluloid) World: On-Screen Representations of Female Characters in the Top 100 Films of 2018. Center for the Study of Women in Television and Film, San Diego State University.

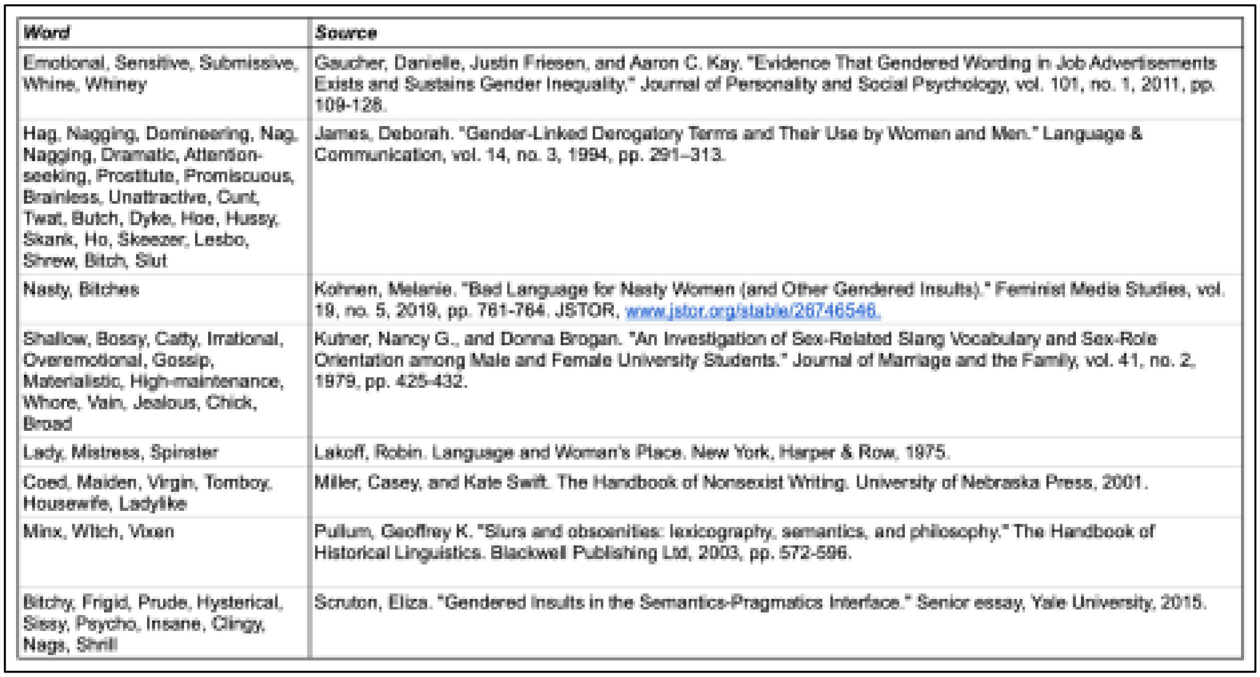
Miller, Casey, and Kate Swift. The Handbook of Nonsexist Writing. University of Nebraska Press, 2001. "Movie Scripts Corpus." Kaggle, 2021, www.kaggle.com/ashishgup/movie-scripts-corpus.

Pullum, Geoffrey K. "Slurs and obscenities: lexicography, semantics, and philosophy." The Handbook of Historical Linguistics. Blackwell Publishing Ltd, 2003, pp. 572-596.

Scruton, Eliza. "Gendered Insults in the Semantics-Pragmatics Interface." Senior essay, Yale University, 2015.

Smith, S., & Choueiti, M. (2019). Inequality in 1,200 Popular Films: Examining Portrayals of Gender, Race/Ethnicity, LGBT, and Disability from 2007-2018. Annenberg Inclusion Initiative, University of Southern California.

"The Movies Dataset." Kaggle, 2021, www.kaggle.com/rounakbanik/the-movies-dataset.



Word	***Source***
Emotional, Sensitive, Submissive, Whine, Whiney	Gaucher, Danielle, Justin Friesen, and Aaron C. Kay. "Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality." Journal of Personality and Social Psychology, vol. 101, no. 1, 2011, pp. 109-128.
Hag, Nagging, Domineering, Nag, Nagging, Dramatic, Attention-seeking, Prostitute, Promiscuous, Brainless, Unattractive, Cunt, Twat, Butch, Dyke, Hoe, Hussy, Skank, Ho, Skeezer, Lesbo, Shrew, Bitch, Slut	James, Deborah. "Gender-Linked Derogatory Terms and Their Use by Women and Men." Language & Communication, vol. 14, no. 3, 1994, pp. 291–313.
Nasty, Bitches	Kohnen, Melanie. "Bad Language for Nasty Women (and Other Gendered Insults)." Feminist Media Studies, vol. 19, no. 5, 2019, pp. 761-764. JSTOR, www.jstor.org/stable/26746546.
Shallow, Bossy, Catty, Irrational, Overemotional, Gossip, Materialistic, High-maintenance, Whore, Vain, Jealous, Chick, Broad	Kutner, Nancy G., and Donna Brogan. "An Investigation of Sex-Related Slang Vocabulary and Sex-Role Orientation among Male and Female University Students." Journal of Marriage and the Family, vol. 41, no. 2, 1979, pp. 425-432.
Lady, Mistress, Spinster	Lakoff, Robin. Language and Woman's Place. New York, Harper & Row, 1975.
Coed, Maiden, Virgin, Tomboy, Housewife, Ladylike	Miller, Casey, and Kate Swift. The Handbook of Nonsexist Writing. University of Nebraska Press, 2001.
Minx, Witch, Vixen	Pullum, Geoffrey K. "Slurs and obscenities: lexicography, semantics, and philosophy." The Handbook of Historical Linguistics. Blackwell Publishing Ltd, 2003, pp. 572-596.
Bitchy, Frigid, Prude, Hysterical, Sissy, Psycho, Insane, Clingy, Nags, Shrill	Scruton, Eliza. "Gendered Insults in the Semantics-Pragmatics Interface." Senior essay, Yale University, 2015.

Negative Word Sourcing