DATA STRUCTURES AND ALGORITHMS

FINAL REPORT

<u>Professor</u>: Kaushik Gopalan

<u>Course code</u>: CSIT206

<u>Topic</u>: Animated Movies Analysis

<u>Contributors</u>: Sneh Puja, Varsha Gunturu, Vineesha Vuppala

Table of Contents-

About the project-

Main content and why

What is our project about?

Our project is based on a dataset consisting of information on animated movies including the financial details, film production and content details and quantitative audience perspectives. We have based our approach on a business management approach, viewing the data as a way to produce output that would give us insights into the considerations required for the future of the production company.

Why did we choose this topic?

The three of us are either minoring in or interested in pursuing Business Analytics, and this is one area we find quite interesting. Moreover, we love watching animated movies and have never explored this sort of data before. Combining these two, we came up with our topic to analyse animated movies.

What information have we found with our code?

1. Top 10 Movie Titles based on Highest Revenue and Highest Popularity
2. Top 10 Production Companies based on Total Revenue
3. Correlations and analysis for different factors
4. Word frequency in the title
5. Word frequency in different years
6. Word frequency in title according to different languages
7. Word frequency vs years
8. Frequencies-years, genres, languages
9. Country Analysis - revenue, IMDb rating, popularity and produced movies

Methodology used

For the top 10s of the different categories, the top 10 movies or production companies were extracted based on the dependent factors of revenue or popularity and ranked. It was based on the descending order of the highest numerical values of revenue or popularity.

For the correlation analysis and regression the various formulas below were used under the cmath library that we used

Correlation coefficient: $r = \frac{n\sum x^2 - (\sum x)^2}{\sqrt{[n\sum y^2 - (\sum y)^2][n\sum xy - \sum x \sum y]}}$

Linear Regression formulas used:

Slope: $\frac{n\sum x^2 - (\sum x)^2}{n\sum xy - \sum x \sum y}$

Intercept: $\frac{\sum y - Slope \times \sum x}{n}$

For the frequency of years, genres and languages

The frequency of years and languages were calculated through the functions countLanguages and countReleaseYearFrequency. The strings were first inputted into a vector and then were arranged in ascending order of alphabet. A loop was used to count and keep track of the frequency, when the loop encountered a different string each time, the count was outputted and then reset to calculate for the next string.

Whereas for genres, each movie contained more than one genre, therefore through the function countAllGenresFrequency, the genres string for each movie was split into each individual genre and then inputted into the vector. The vector was then arranged in alphabetical order thereby grouping the frequency together in the vector. A loop was used again here to count the frequency of each genre group arranged and when the genre changes the count is outputted and then reset to calculate for the next genre word.

For the frequency of the words in the title

A. General Word Frequency Analysis

   The word frequency is stored in a structure called WordFrequency which stores the word and the frequency of it.

   Each function takes in a vector of movies (a custom class we created to store the attributes for the movies) as well as a vector of strings which are the words to ignore while doing the analysis as they are not of relevance to understanding the themes.

   The function iterates through every movie in the vector of movies. It converts the title to string steam so each word of the title can be extracted through the stream extraction operator >> into the variable word. Then each letter of the word is explicitly converted to lowercase.

Then after the word is converted to its lowercase form, a binary search is performed to check whether the word needs to be ignored or not.

Only if the word is not found in the ignoredWords vector, does the function count the frequency.

First it checks whether this is the first time the word has been come across in the titles. If it has already been encountered, it simply increments the current frequency. If it is the first time, it updates the frequency to one.

B. Word Frequency Analysis by Language
   It is done by the function countTitleWordsByOriginalLanguage which returns a vector of vector of word frequencies. This is done to have separate vectors of word frequencies for every language. For each vector of word frequencies, the first word frequency includes the name of the language, thus helping group the words of different languages. In addition to the general parameters as discussed above, it also includes a vector of strings for languages to ignore (as an insignificant number of movies in this particular dataset are in that particular language) . The general process is the same.

C. Word Frequency Analysis by Year
   It is done by the function countTitleWordsByYear which returns a vector of pairs of an integer(for year) and a vector of word frequencies. This is done to have separate vectors of word frequencies for every year. It defines a vector of years to be excluded within the function itself. The release dates are tokenized manually to get the year out of the date which is then converted to an integer type. The general process is the same.

Merge sort was used for the majority of the sorting. Only bucket sort was used to sort the ignoredLanguages vector.

Findings and Analysis

A. Ranking Analytics

Findings-

The data we have found:
Top 10 Movie Titles with Highest Revenue and Their Production Companies:

1. Title: Frozen II - Revenue: $1450026933
   Production Company: Walt Disney Pictures, Walt Disney Animation Studios,

2. Title: The Super Mario Bros. Movie - Revenue: $1355725263
   Production Company: Universal Pictures, Illumination, Nintendo,

3. Title: Frozen - Revenue: $1274219009
   Production Company: Walt Disney Pictures, Walt Disney Animation Studios,

4. Title: Incredibles 2 - Revenue: $1242805359
   Production Company: Walt Disney Pictures, Pixar,

5. Title: Minions - Revenue: $1156730962
   Production Company: Illumination, Universal Pictures,

6. Title: Toy Story 4 - Revenue: $1073394593
   Production Company: Walt Disney Pictures, Pixar,

7. Title: Toy Story 3 - Revenue: $1066969703
   Production Company: Pixar, Walt Disney Pictures,

8. Title: Despicable Me 3 - Revenue: $1031552585
   Production Company: Illumination, Universal Pictures,

9. Title: Finding Dory - Revenue: $1028570889
   Production Company: Pixar, Walt Disney Pictures,

10. Title: Zootopia - Revenue: $1023784195

Production Company: Walt Disney Animation Studios, Walt Disney Pictures,

Top 10 Movie Titles with Highest Popularity and Their Production Companies:

1. Title: Elemental - Popularity: $1008.94

Production Company: Walt Disney Pictures, Pixar,

2. Title: Carl's Date - Popularity: $819.429
   Production Company: Pixar,

3. Title: Spider-Man: Across the Spider-Verse - Popularity: $512.336
   Production Company: Columbia Pictures, Sony Pictures Animation, Lord Miller, Pascal Pictures, Arad Productions,

4. Title: Teenage Mutant Ninja Turtles: Mutant Mayhem - Popularity: $484.876
   Production Company: Paramount, Nickelodeon Movies, Point Grey Pictures, Mikros Animation, Cinesite Animation,

5. Title: Scooby-Doo! And Krypto, Too! - Popularity: $411.09
   Production Company: Warner Bros. Animation,

6. Title: The Super Mario Bros. Movie - Popularity: $410.411
   Production Company: Universal Pictures, Illumination, Nintendo,

7. Title: Ruby Gillman, Teenage Kraken - Popularity: $321.777
   Production Company: Universal Pictures, DreamWorks Animation, dentsu,

8. Title: Epic Tails - Popularity: $304.93
   Production Company: TAT Productions, Apollo Films,

9. Title: One Piece Film Red - Popularity: $293.869
   Production Company: Toei Animation, Shueisha, Toei Company, Fuji Television Network, Bandai, Bandai Namco Entertainment, ADK Emotions, dentsu,

10. Title: Resident Evil: Death Island - Popularity: $264.731
    Production Company: CAPCOM, TMS Entertainment, Quebico, UNLIMITED PRODUCE by TMS, Sony Pictures, Sammy, Robot Communications,

<u>Top 10 Production Companies by Total Revenue:</u>

1. Pixar, Walt Disney Pictures - Total Revenue: $7934987635, Produced Countries: United States of America
2. Illumination, Universal Pictures - Total Revenue: $4752136567, Produced Countries: United States of America
3. Walt Disney Pictures, Walt Disney Animation Studios - Total Revenue: $4555772760, Produced Countries: United States of America

4. DreamWorks Animation, Paramount - Total Revenue: $4131753443, Produced Countries: United States of America
5. Blue Sky Studios, 20th Century Fox Animation, 20th Century Fox - Total Revenue: $3934506069, Produced Countries: United States of America
6. Walt Disney Pictures, Walt Disney Feature Animation - Total Revenue: $3902741165, Produced Countries: United States of America
7. Pixar - Total Revenue: $3863299965, Produced Countries: United States of America, Canada, United States of America
8. Walt Disney Pictures, Pixar - Total Revenue: $3639185553, Produced Countries: United States of America
9. Walt Disney Animation Studios, Walt Disney Pictures - Total Revenue: $2664908810, Produced Countries: United States of America
10. DreamWorks Animation, 20th Century Fox - Total Revenue: $2238904436, Produced Countries: United States of America

Findings-

1. The top-grossing movies and their production companies are in the top 10 for both
2. A production company with a track record of successful films is more likely to produce another successful movie
3. Shows consistency of quality movies by a company
4. Investors and stakeholders may use this information to identify production companies with a strong track record of profitability

Top 5 countries with the highest revenue:

1. United States of America
2. Japan
3. China
4. Japan, United States of America
5. Australia, United States of America

Top 5 countries with the highest IMDb rating:

1. United States of America
2. Japan
3. France
4. Soviet Union
5. United Kingdom

Top 5 countries with the highest popularity:

1. United States of America
2. Japan
3. France
4. United Kingdom

5. Canada

Top 5 countries with the highest total number of movies produced:
1. Kazakhstan, Russia, Turkey, Uzbekistan
2. Kazakhstan
3. Netherlands, Syrian Arab Republic, Turkey
4. Austria, Canada
5. China, Portugal

Analysis-
1. The United States of America consistently emerges as the top contributor to revenue, underscoring its dominance in the global animated movie market. Major production companies like Walt Disney Pictures and Pixar play significant roles in driving revenue.
2. The United States of America and Japan lead in IMDb ratings, reflecting the quality and popularity of their animated movies. France also stands out for its high-quality productions, contributing to its recognition in the industry.
3. The United States of America and Japan dominate in terms of popularity, showcasing the widespread appeal of their animated movies, indicating the global reach of their productions.
4. The dataset includes a mix of countries with varying levels of animation production. Kazakhstan appears multiple times in the top five, suggesting a significant presence in animated movie production. Other countries like the Netherlands, Turkey, and Canada also feature, highlighting the diversity of global animation hubs.

Countries and their analysis with the Top 10s
1. United States of America (USA):
    - The USA emerges as the dominant player across all metrics, including revenue, IMDb ratings, popularity, and total number of movies produced
    - Major production companies such as Walt Disney Pictures, Pixar, Universal Pictures, and Paramount contribute significantly to the industry's success
    - The USA's strong presence underscores its status as a powerhouse in the global animated movie market, driven by established studios and widespread audience appeal
2. Japan:
    - Japan ranks consistently high in terms of revenue, IMDb ratings, and popularity, indicating a strong and diverse animated movie industry
    - Notable titles such as "One Piece Film Red" and "Spider-Man: Across the Spider-Verse" contribute to Japan's success in both domestic and international markets
    - The country's unique animation style and storytelling techniques continue to captivate audiences worldwide, cementing its position as a key player in the industry
3. China:
    - China emerges as a significant contributor to revenue, with notable titles like "Resident Evil: Death Island" generating substantial popularity

- Despite not being as prominent in IMDb ratings or popularity, China's growing influence in the global market is evident through its collaboration with international studios and investment in animated movie production

4. France:
    - France showcases a strong presence in terms of IMDb ratings, indicating a focus on quality and critical acclaim in its animated movie productions
    - While not as prominent in revenue or popularity compared to the USA or Japan, France's emphasis on artistic expression and storytelling contributes to its recognition in the industry
5. United Kingdom (UK):
    - The UK ranks consistently among the top countries in terms of IMDb ratings and popularity, reflecting its contribution to the global animated movie landscape
    - Titles like "Carl's Date" and "Scooby-Doo! And Krypto, Too!" highlight the UK's diverse storytelling and animation styles, attracting audiences worldwide

B. Correlation and Regression

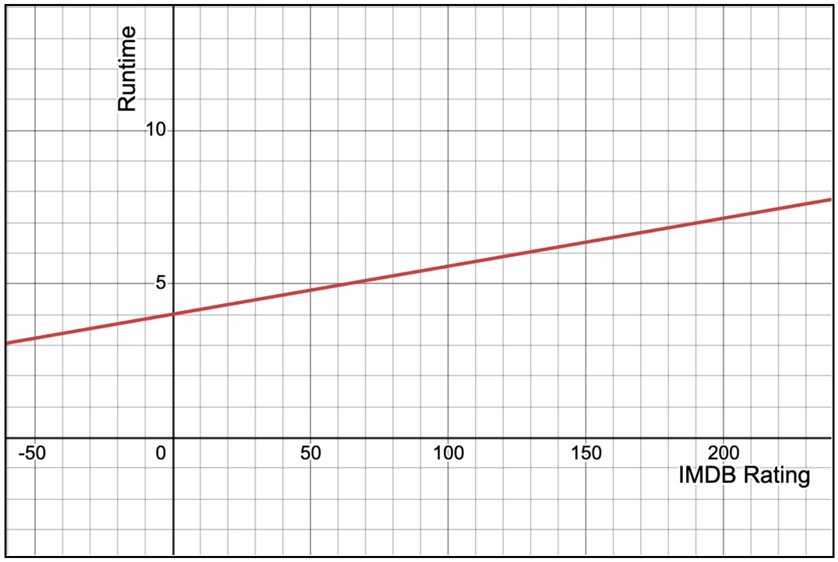
Findings-

The data we have found:

Features	**r**	**R^2**	**Regression Equation**
Runtime and IMDb rating	0.211871	0.04489	y = 0.0156118x + 4.01766
Runtime and Popularity	0.18074	0.03267	y = 0.0512528x + 1.40063
Runtime and Revenue	0.12474	0.01556	y = 111537x + 266138
Revenue and Popularity	0.386493	0.14938	y = 1.22579e-07x + 2.34214
Revenue and IMDB_rating	0.0640293	0.00410	y = 5.27684e-09x + 4.40567
IMDB_rating and Popularity	0.126751	0.01607	y = 0.487792x + 0.571982
Budget and Revenue	0.775211	0.60095	y = 3.1396x + 163810
Budget and Runtime	0.149106	0.02223	y = 6.75401e-07x + 25.2771
Budget and Popularity	0.380636	0.14488	y = 4.88923e-07x + 2.26299
Budget and IMDB_rating	0.0701381	0.00492	y = 2.34101e-08x + 4.40001

Graphs to visualise the data -

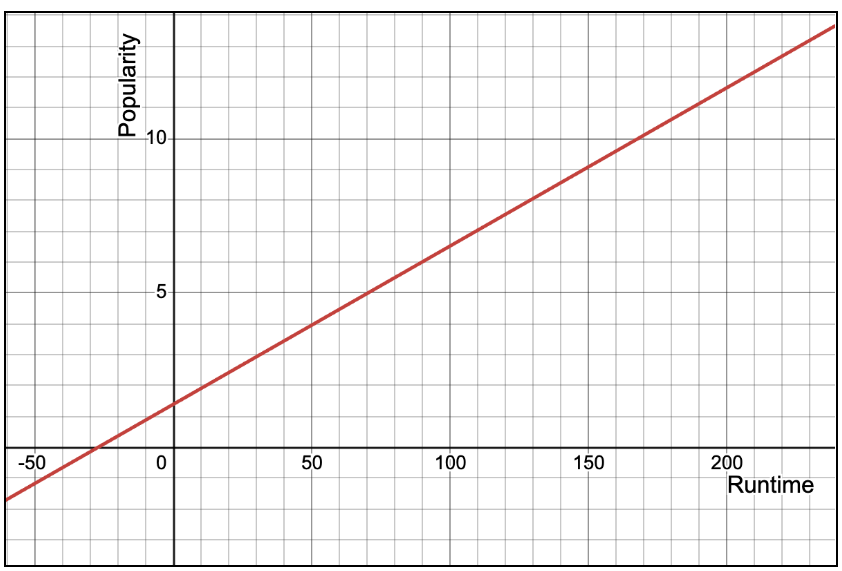
Runtime and IMDb rating:



Correlation Coefficient: 0.211871(weak positive correlation)
R^2: 0.04489
Linear Regression Equation: y = 0.0156118x + 4.01766

The above correlation coefficient shows that longer movies might slightly tend to have higher ratings. The R square is 0.04489, meaning that only 4.4% of variation in y is explained by variation in x, which is also an indicator that the regression model cannot accurately predict the IMDB rating based on runtime alone.

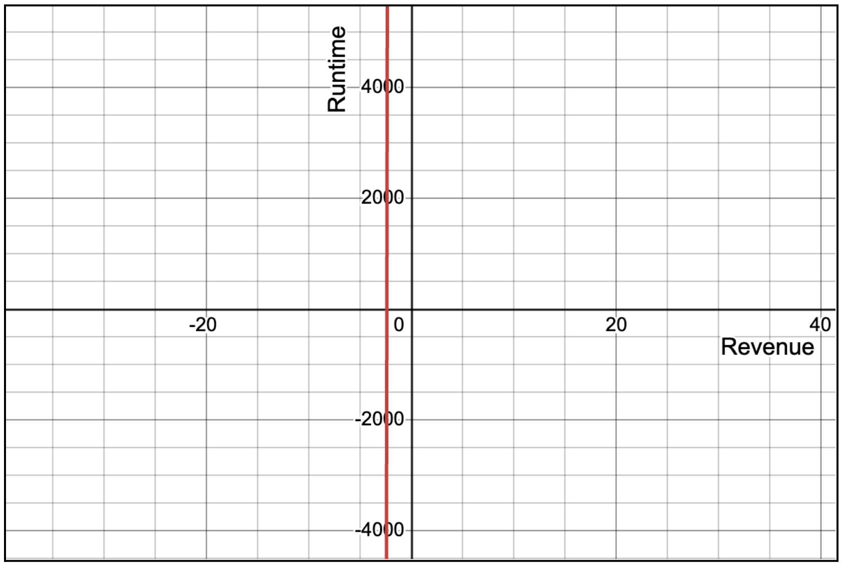
Runtime and Popularity:



Correlation Coefficient: 0.18074 (weak positive correlation)
R^2: 0.0326669476
Linear Regression Equation: y = 0.0512528x + 1.40063
The above correlation coefficient shows that longer movies may attract slightly more attention, but the correlation is not strong as it is only 18%. The R square of the regression which is only

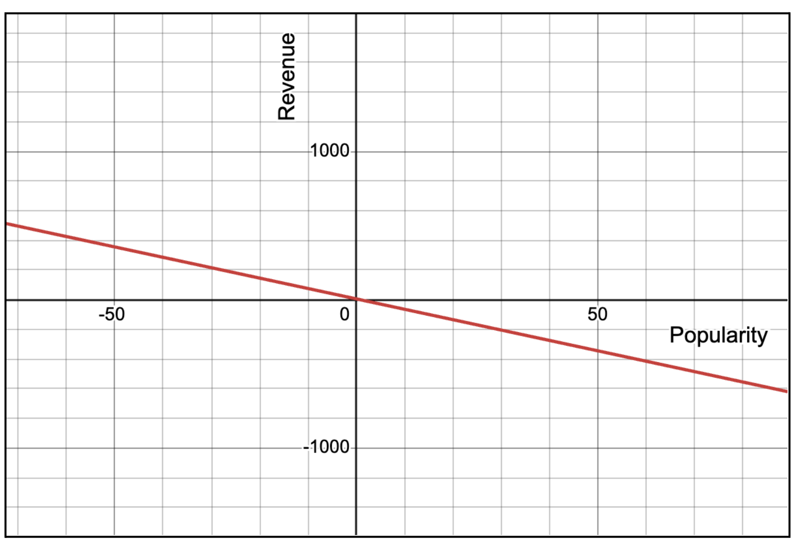
3.2% shows that the linear regression model is a weak fit and that only 3.2% variation in popularity is explained by variation in runtime.

Runtime and Revenue:



Correlation Coefficient: 0.12474 (very weak positive correlation)
R^2: 0.01556
Linear Regression Equation: y = 111537x + 266138
Runtime and revenue are 12.4% correlation, which provides evidence that the length of a movie doesn't significantly impact its revenue. The same can also be seen in the linear regression, here according to R square only 1.5% of variation in runtime is explained by variation in revenue.

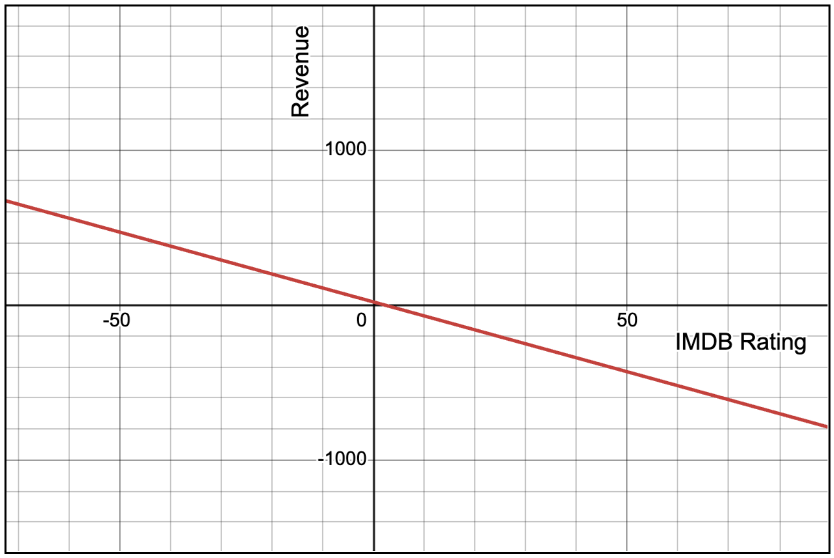
Revenue and Popularity:



Correlation Coefficient: 0.386493 (moderate positive correlation)
R^2: 0.14938
Linear Regression Equation: y = 1.22579e-07x + 2.34214

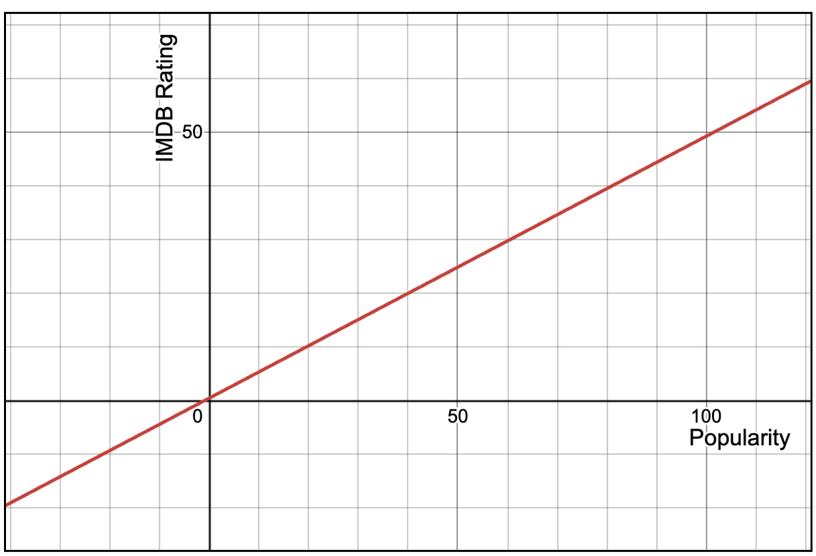
Revenue and popularity are 38% correlated which shows that higher revenue movies tend to be more popular, which is a reasonably expected relationship. Whereas the r square of regression shows that only 14.9% of variation in popularity is explained by variation in revenue, therefore the model is a low moderate fit.

Revenue and IMDB_rating:



Correlation Coefficient: 0.0640293 (very weak positive correlation)
R^2: 0.00410
Linear Regression Equation: y = 5.27684e-09x + 4.40567
Revenue and IMDB rating are 6.4% correlated that shows that high revenue doesn't necessarily translate to high IMDb ratings, and vice versa. The r square of regression is also 0.4% which is very low and shows that the regression model is not a good fit.

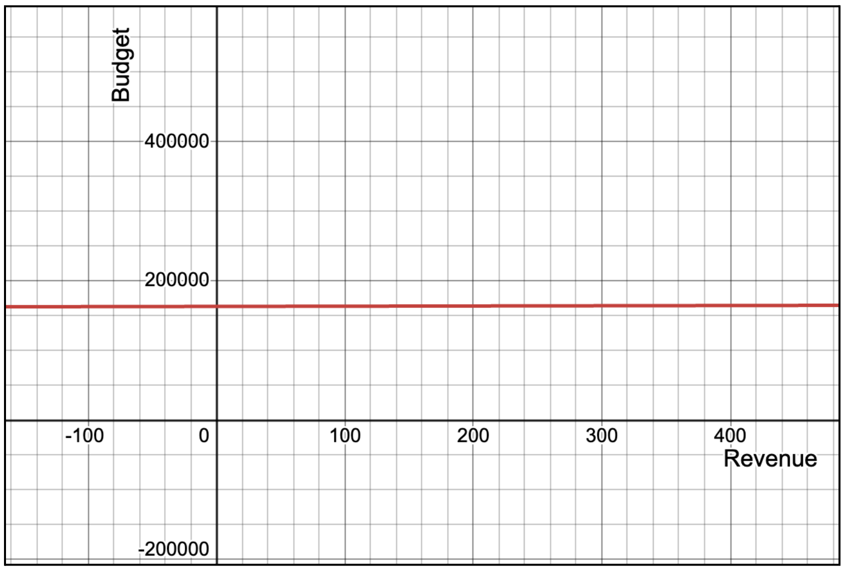
IMDB_rating and Popularity:



Correlation Coefficient: 0.126751(very weak positive correlation)

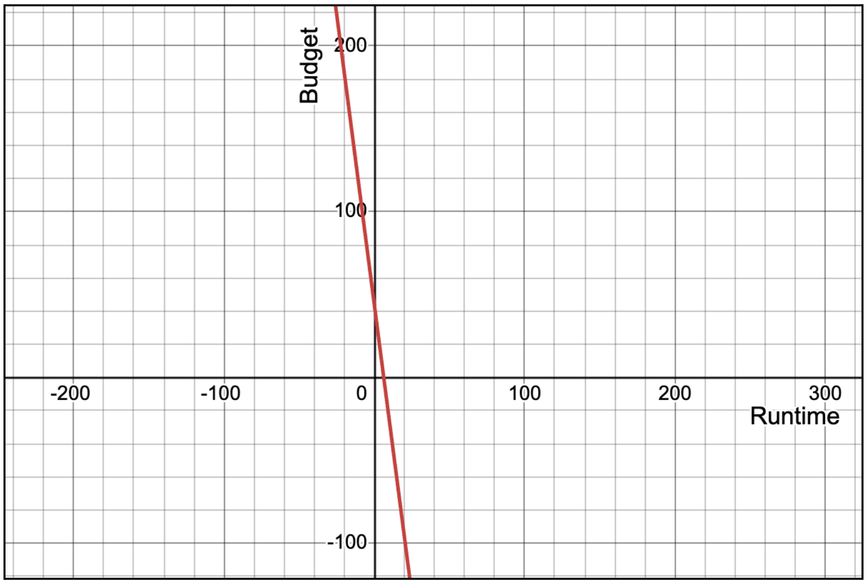
R^2: 0.01607
Linear Regression Equation: y = 0.487792x + 0.571982
IMDB rating and popularity are only 12.6% correlated, this shows that there's not a strong relationship between how well-rated a movie is on IMDb and how popular it is. The same can be seen in the regression model where r square is only 1.6%, showing that the model is poor fit.

Budget and Revenue:



Correlation Coefficient: 0.775211 (strong positive correlation)
R^2: 0.60095
Linear Regression Equation: y = 3.1396x + 163810
Budget and revenue are 77.5% correlates showing that movies with higher budgets tend to generate higher revenues. The r square is also 60% meaning that 60% variation in revenue is explained in variation in budget. This linear model is a strong fit.

Budget and Runtime:



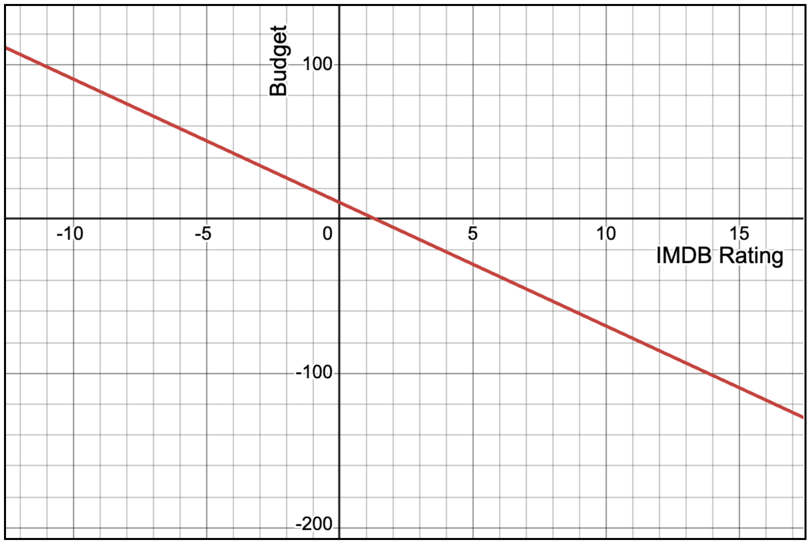
Correlation Coefficient: 0.149106 (weak positive correlation)
R^2: 0.02223
Linear Regression Equation: y = 6.75401e-07x + 25.2771
Budget and runtime have a 14.9% correlation which means that the budget doesn't heavily influence the length of a movie. Quite contrary to beliefs that more the budget means a longer runtime, the correlation coefficient shows only 14.9%. The linear model with 2.2% r square also shows to be a poor fit.

Budget and Popularity:

Correlation Coefficient: 0.380636 (moderate positive correlation)
R^2: 0.14488
Linear Regression Equation: y = 4.88923e-07x + 2.26299

Budget and popularity are 38% correlated showing that movies with higher budgets tend to be more popular, although other factors also play a role in a movie's popularity. The r square is 14.4%, which shows that only 14.4% of variation in popularity is explained by variation in budget.

Budget and IMDB_rating:



Correlation Coefficient: 0.0701381 (very weak positive correlation)
R^2: 0.00492
Linear Regression Equation: y = 2.34101e-08x + 4.40001
Budget and IMDB are 7% correlated, which shows that budget alone doesn't determine how well a movie is rated on IMDb. The same can be seen in the linear regression with only 0.4% r square.
Analysis-

Correlations	Analysis
Correlation between runtime and IMDb rating: 0.211871	• Weak positive correlation • Audiences may perceive longer movies as more substantial or engaging
Correlation between runtime and popularity: 0.18074	• Weak positive correlation • Longer movies may attract more attention from audiences
Correlation between imdb_rating and popularity: 0.126751	• Weak positive correlation • A modest overlap between critical acclaim and audience appeal.
Correlation between revenue and popularity: 0.386493	• Moderate positive correlation • Movies with higher revenues --> more popular among audiences

	• Financial success and audience appeal go hand in hand
Correlation between revenue and imdb_rating: 0.0640293	• Very weak correlation • Movie's box office performance does not necessarily reflect its critical acclaim or quality
Correlation between runtime and revenue: 0.124748	• Weak positive correlation • Factors other than runtime play a more significant role in determining a movie's revenue
Correlation between budget and revenue: 0.775211	• Strong positive correlation • Movies with higher budgets tend to generate significantly higher revenues
Correlation between budget and runtime: 0.149106	• Weak positive correlation • Budget constraints may not heavily influence the duration of a film
Correlation between budget and popularity: 0.380636	• Moderate positive correlation • Higher budgets are generally more popular among audiences, possibly due to increased marketing efforts or higher production values
Correlation between budget and imdb_ratings: 0.0701381	• Very weak correlation • Movie's budget does not necessarily guarantee a higher IMDb rating or critical success

C. Frequencies

Years

1. 2021: 1942
2. 2022: 1351
3. 2023: 1250
4. 2020: 1067
5. 2019: 920
6. 2017: 810
7. 2018: 784
8. 2015: 778
9. 2016: 756
10. 2014: 756
11. 2013: 734
12. 2012: 668
13. 2010: 618

14. 2011: 616
15. 2009: 533
16. 2008: 494
17. 2006: 484
18. 2007: 461
19. 2005: 445

The above are the years with most animated movies released. We can see that 2021 had the most animated movies ever released and that overall, the number of movies have been increasing significantly over the years with the number of movies tripling between 2011 and 2021, mostly attributed to the biggest jump from 2020 to 2021 with the number of released movies almost doubling. A possible reason could be the movies that were supposed to release in 2020 got delayed due to Covid-19 and had to be released the next year. However, proceeding that, there was a sharp dip in movies released in 2022 and 2023.

Genres:

1. Animation: 30488
2. Comedy: 6060
3. Family: 5419
4. Fantasy: 2989
5. Adventure: 2819
6. Drama: 2003
7. Action: 2040
8. Science Fiction: 1984
9. Music: 1121 (tie)
10. Documentary: 943 (tie)
11. Horror: 943 (tie)
12. Mystery: 434
13. Romance: 647
14. Crime: 264
15. History: 314
16. War: 266
17. Thriller: 240
18. Western: 124

The above shows the most common genres released through all the years in the dataset. We can see that the most common genres other than animation are comedy, family, fantasy and adventure, and drama which are more lighthearted compared to the bottom 10 which include horror, mystery, crime, war, thriller, and western.

Languages:

Top 10 Languages

1. English (en): 16951 movies
2. Japanese (ja): 3714 movies
3. French (fr): 2039 movies
4. Russian (ru): 1391 movies

5. Spanish (es): 872 movies
6. German (de): 619 movies
7. Portuguese (pt): 459 movies
8. Chinese (zh): 550 movies
9. Italian (it): 289 movies
10. Czech (cs): 295 movies

The above shows the top 10 languages with most animated movies released. English dominates as its one of the languages which holds the most cultural soft power. Following that is Japanese, which is not surprising as anime, a specific animated form of Japanese animation, has become widely popular across countries and cultures in the past decade. Given that France has one of the biggest movie production companies and has had a deep influence on cinema itself since its origin, it comes third. Though the origins of Russian animation date back to the early 1990s there was a significant increase in animated movies since the time of the Cold War to compete with the U.S for soft power. Now it provides alternatives to Disney and other major U.S as well as Japan animation companies since it does not have good political relations with both. We see that though it has a significant number of movies, it is not amongst the top 5 popular countries. Spanish Animation as well has had a steady rise in popularity, particularly in the last decade.

D. Thematic Analysis Through Words in Movie Titles

In General

Most of the words depict innocence and family/children oriented words as majority of these movies are for children. In addition, the words are simple to understand and usually short in size so that the target audience can understand the title. The most common word is little, explicitly addressing the target audience. In contrast, another common word is big, one reason could be that they are very commonly used words by children and also children generally think in like antonyms or opposites. The second most common word is christmas. This may be an important theme as the concept of Christmas is very family oriented and it's a tradition for families in Western countries to watch movies together since the entire family usually has vacation or days off. The third most common word is love, weaving into the common theme. There are some more fantasy/adventure themes that are frequently used in titles displayed in words such as adventure, legend, tale, magic, dragon, and king. There are also a few animals - cat, mouse (this could be in regards to 'mickey mouse'), and dog as well as colours pink and blue.

Languages

English

Innocent child-like family-oriented themes are the most relevant. Christmas is the most popular word, followed by little. The most common colour is pink.

Chinese

Very legend/ old tale related .

Japanese

More fantasy/mystical themes are common with words like legend, dragon, adventure, and world. Characters like doraemon are also common. There are many stories centred around girl characters.

French

In contrast to Japanese, there are more stories centred around men. Little is the most common word. Magic and night themes are common.

Russian

The common themes are adventures and tales. Details of the landscapes come in with the common words like snow, wolf, and cat.

Italian

Mostly fantasy themes. Lots of specific characters and stories

Years

1800s

Words like jeu (French for game), clown, musician, and pierrot (a comedic pantomime character in French and Italian cinema allude to entertainment and performance.

1900s-1910

Words like enchanted, villain, haunted, ghosts, nightmare, lightning, fantasmagorie (French for phantasmagoria - a sequence of real or imaginary images like that seen in a dream), dreams allude to more mystical dark complex stories.

1910s-1920

Words like esprits(French for spirit), vengeance, fantasies, revenge, dreams, ghost, robbery, ouija, follow the dark team for the 1900s to 1910s but there's also an increase in poetic themes with words like beautiful, portrait, poses, art, kiss, moon. An influence of World War 1 is seen with words like colonel, war, captain, and lusitania (a ship that was torpedoed by a German U-boat that became propaganda for the British and Americans).We also see a usage of highly specific characters and names such as charlie, bobby, grogg, momi, heeza, joneses, leukanida, nemo, and rosalie.

1920-1930

Again specific characters and names are common with words such as cinderella, jack, felix, mickey, and alice. A lot of animals are also included in titles such as cat, rat, monkey, horses, grasshopper, tail, and dog. This is a time of whimsical classic stories and fables that movies with titles regarding very specific stories, or tales.

1930-1940

This is a time where there were many mickey mouse movies as well as the character/tale-specific stories. The word little also became very common in this decade.

1940-50

Animals are the most common theme with fox, mouse, rabbit, dog, bunny, wolf, hare, mouse, duck, cats, bugs, mice, chick, and lion.

1950-60

The theme of animals and the word little continues.
<u>1960-1970</u>
Animals yet again are common, as well as some more family oriented and celebration themed words such as love, christmas, birthday. The word pink becomes very common as well as little.
<u>1970-1980</u>
The popularity of pink continues. There are less objects and circus/entertainment themes. Specificity in characters and names remains.
<u>1980-1990</u>
Christmas increases in popularity. Animals and specificity are still prevalent. Childish adventure themes are present.
<u>1990-2000</u>
Adventure and fantasy themes increase with words such as adventure, prince, dragon, hunter, world, cowboys, secret, legend, stories, king, galaxy, legend, time, hercules. Christmas is the most popular word.
<u>2000-2010</u>
Interestingly science and science-fiction becomes a common theme, along with Christmas, adventures, dogs, cats, little, legend, and friends. There are again specific characters but the types of characters and their stories have changed and there are more generalised themes as well.
<u>2010-2020 onwards</u>
Love becomes a popular word, along with little, friends, happy, christmas, home. Superhero and fantasy themes are also relevant.

Night, stars, moon, day consistently occur throughout the years and languages. There are a lot of contrasts like these - night and day, sun and moon, big and little, blue and pink,etc. The themes shift from more whimsical, theatrical and playful, to more modern takes on entertainment. There is an increasing generalisation of titles as well though certain specific characters are still common. There are some darker themes but they are not quite common and more lighter mystery, adventure, fantasy and superhero/action themes are common.

Applications of our project in real-life-

- Production companies can use insights to decide which animated movies to produce and how to allocate resources effectively

- Marketing teams can tailor campaigns based on audience preferences, genres, and popular regions, popular themes and content (according to region as well)
-
- Audience engagement → Streaming platforms can use data to offer personalized recommendations, enhancing user satisfaction

- Animators can identify trends and preferences to guide the creation of new animated content

- Financial Investment → Investors can assess the profitability of animated movie projects using predictive models

- Distribution companies can optimize movie releases based on regional demand and localization opportunities

Here is an article that matched our findings on genres and its frequencies - https://www.linkedin.com/advice/0/what-most-common-themes-animation-across-different-genres-qaaxe#:~:text=Drama%20and%20romance%20animation%20often,%2C%20Toy%20Story%2C%20and%20Up.

Here are some articles where we found that are findings matched with the Top 10s
1. https://academyofanimatedart.com/animation-market-statistics/ → top grossing movie being Frozen II
2. https://vitrina.ai/blog/discover-the-best-animation-studios-worldwide-top-10-animation-companies/ → top animation companies names like Pixar, Walt Disney Pictures and more

Limitations-

1. Inaccurate production company names:
    - The dataset contains inconsistencies in naming conventions for production companies, leading to difficulties in accurately identifying and analyzing their contributions to animated movies

2. Repeated countries for production companies:
    - Some production companies are associated with multiple countries, resulting in duplicate entries for the same company and potentially skewing analyses based on country-specific data

3. Handling null values and non-alphanumeric data:

- Null values are present across various columns in the dataset, requiring proper handling to avoid errors in analysis
- Additionally, some rows contain non-alphanumeric data, such as symbols, which may disrupt data processing and analysis if not addressed appropriately

4. Variability in date formats:
    - The representation of release dates within the dataset is inconsistent, with dates appearing in multiple formats. This variability complicates the extraction and interpretation of date-related information.

5. Function complexity and data type variability:
    - In the functions of the code, each individual function was dealing with different pairs of data types, hence a multitude of sort function were needed for the different parameters

Future scope-

1. Seasonality: Look into how production companies plan out their movies, if there is a theme for certain movies at certain time frames (eg: Christmas movies during end of the year)

2. Time frames: Look into how production companies release their movies. Which are the popular months and why

3. Improved data handling: Make the process of cleaning and organizing data smoother to ensure better accuracy and reliability in our findings

4. Smart predictions: Use smarter algorithms to predict which animated movies will do well financially and which ones will be popular with audiences

5. Stay up-to-date: Create a system that keeps track of movie trends in real-time, so we always have the latest information for decision-making

6. Recommendations: Build systems that suggest movies you might like based on what you've watched before, making it easier to find new favorites