# *MOVIE ANALYSIS*

# *FIRST PHASE PROJECT*

*BY KINYANJUI KAMAU CHRIS.*

Kinyanjui Kamau Chris. 27th Aug 2022.

# Table of Contents

Kinyanjui Kamau Chris. 27th Aug 2022.

# Table of Figures

Kinyanjui Kamau Chris. 27th Aug 2022.

# *MOVIES ANALYSIS PROJECT.*

## *1. BUSINESS UNDERSTANDING*

### UNDERSTANDING THE PROBLEM

I have assumed the role of a Data Scientist hired by Microsoft. My task is to use the six provided datasets from the various Movie Database website found online in order to give them actionable insights on how to go about creating a new movie studio.

- This will be done by giving an insight on the popular languages and whether to focus on Domestic market or World-wide market.
- What is the best time of the month to release movies and the preferred length or movie length/duration
- The most popular genres.

### PROBLEM STATEMENT

I am to explore the types of films doing the best on the box office.

My hypothesis will be ;

- The International market will be more profitable than the domestic.
- The English language is by far the most available language for movies.
- There are consistent times of the month every year which yields lower profits when movies are released on, termed as 'Dump Months' and also there is a preferred length of time for movies which the audience can make time.
- There are genre types which are preferred by the majority of movie watchers.

## 2. DATA UNDERSTANDING.

## DATA COLLECTION

The data was collected from github . There are 6 datasets provided.

### *1. "imdb.sql" database.*

Provided by the **IMDb** (**Internet Movie Database**). This is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDb began as a fan-operated movie database on the Usenet group "rec.arts.movies" in 1990, and moved to the Web in 1993. It is now owned and operated by **IMDb.com, Inc.**, a subsidiary of Amazon. More info on them can be gotten from here.

### *2. "tmbd.csv" database.*

Provided by The Movie Database (TMDB). This is a *community built* movie and TV database. Every piece of data has been added it's community . More information on them here.

### *3.* **"bom.movie_gross.csv"** *database.*

This was supplied by the Box Office Mojo {bom}. **Box Office Mojo** is an American website that tracks box-office revenue's in a systematic, algorithmic way. The site was founded in 1998 by Brandon Gray, and was bought in 2008 by IMDb, which itself is owned by Amazon.

More information on them here.

### *4. rt.movie_info.tsv.gz and rt.reviews.tsv.gz database.*

This database was provided from the Rotten Tomatoes website. Rotten Tomatoes and the Tomato-meter score are trusted recommendation resources for quality entertainment for a lot of people on the internet. They provide fans with a comprehensive guide to what's Fresh – and what's Rotten.

### *5. tn.movie_budgets.csv.gz database.*

This database was provided by The Numbers website. They are website which tracks movie information as they continuously update their data.

# DATA DESCRIPTION

### *1. "imdb.sql" database.*

| Tables | Description |
|---|---|
| movie_basics | A table with the information of various movies like start year, runtime and genres linking with their respective movie_id. |
| directors | A table linking directors to their movies through the movie id. |
| known_for | A table that links persons to the respective films they are known for. |
| movie_akas | A table with movies and regions with their respective language and original title. |
| movie_ratings | Contains the average ratings and number of votes for each movie represented by their respective movie ids |
| persons | A table with the name and person id of various personal. It also includes their birth_year and death_year together with their primary professions. |
| principals | A Table of key individuals detailing their jobs, category,characters from movies and movie id. |
| writers | Table of screen play writers and their respective films as movie id |

*Figure 1: imdb.sql table*

### *2. "tmbd.movies.csv" database.*

| **Columns** | **Description** |
|---|---|
| genre_ids | A column of genres in number form each a key their respective value of genre .Eg<br>99 = 'documentaries' |
| id | The unique id from The Movie database website system. |
| original_language | The original language for the movie record. |
| original_title | The original title of the movie, this means if it was a Russian movie the the original title is its Russian title and not its translated title. Eg ***Tres metros sobre el cielo*** (Spanish)which under title is translated to ***Three Steps Above Heaven.*** |
| popularity | A float value which represents the movies popularity based on an algorithm which takes into account its daily rating, number of votes and its past critics as it updates continually. |
| release_date | A string of dates which represent when the recorded movie was released. |
| title | The english version of the movies title. |
| vote_average | Its average rated score, highest being 10 lowest being 1. |
| vote_count | No of people who have voted on the recorded movie. |

*Figure 2: tmbd.movies.csv*

### *3.* "bom.movie_gross.csv" *database.*

| **Columns** | **Description** |
|---|---|

| | |
|---|---|
| title | Name of the recorded movie. |
| Studio | Abbreviation of the Studio that worked on the recorded movie. |
| Domestic_gross | The gross amount of money the recorded film made locally. |
| Foreign_gross | The gross amount of money made by the recorded film Internationally. |
| year | Year the recorded movie was released. |

*Figure 3: bom.movie_gross.csv*

### 6. rt.movie_info.tsv.gz

| Column | Description |
|---|---|
| Synopsis | A brief description of the redcorded movie. |
| rating | A rating for the recorded movie that represents the age of the target audiences, the unique ratings are,'**R**', '**NR**', '**PG**', '**PG-13**', '**G**', '**NC17**'. |
| genre | A list of the type of movie in the recorded. |
| director | A record that shows the director of the movie. |
| writer | A record that provides the name of the movies writer. |
| theater_date | The date which the recorded film hit theaters. |
| dvd_date | The date which the recorded movies were available on DVD. |
| currency | The currency of the amount of money the movie made in the BOX Office |
| Box Office | Money earned in the box office by the movie recorded. |
| runtime | Length of the recorded movie. |

| | |
|---|---|
| studio | Studio in charge of the recorded movie. |

*Figure 4: rt.movie_info.tsv.gz*

### *7. rt.reviews.tsv.gz database.*

| **Column** | **Description.** |
|---|---|
| review | A record of a critics review for the particular movie. |
| rating | The rating of the movie out of the specified value, either 5 others are out of 10. not a constant method. |
| fresh | A rating of the movie but only between two values, 'fresh' for good or 'rotten' for bad. |
| critic | Name of the person who left the review on the record. |
| top_critic | A record for a boolean of whether the critic was the top(1) or not (0). |
| publisher | Name of the person or for some records the website that left the critic. |
| date | Date the record was made. |

*Figure 5: rt.reviews.tsv.gz*

### *8. tn.movie_budgets.csv.gz database.*

| **Column** | **Description** |
|---|---|
| Movie | Name of the recorded film. |
| production_budget | Amount of money spent to make the recorded movie. |
| domestic_gross | Amount of money the recorded movie made locally. |
| release_date | The date the movie came out. |
| worldwide_gross | Amount of money the film made Internationally. |

*Figure 6: tn.movie_budgets.csv.gz*

# *DESCRIBING THE QUESTION*

As a hired Data Scientist, I am to find the types of films which are currently doing the best and translate them into actionable insights.

### DEFINING THE METRIC OF SUCCESS

When I show through visualizations that;

- There exist the worst months to release films.
- There are preferred genres.
- More money is to be made Internationally.

# *3. DATA PREPARATION*

### SELECTING DATA

I will select the following tables to work with;

1. tn.movie_budgets.csv.gz database.

Chosen because it has the production_budget which can be used to create two new columns of domestic_profits and worldwide_profits.

It also has the release data which can be useful.

2. tmbd.movies.csv database.

I selected this datasets because of the original_language, popularity, vote_average and the genres_id columns.

The table below represents the genre keys from their website.

*Figure 7: Genre key ID for the TMbd data*

| **MOVIE Genre** | **Genre ID** |
|---|---|
| Action | 28 |
| Adventure | 12 |
| Animation | 16 |
| Comedy | 35 |
| Crime | 80 |
| Documentary | 99 |
| Drama | 18 |
| Family | 10751 |
| Fantasy | 14 |
| History | 36 |
| Horror | 27 |
| Music | 10402 |
| Mystery | 9648 |
| Romance | 10749 |
| Science Fiction | 878 |
| TV Movie | 10770 |
| Thriller | 53 |
| War | 10752 |
| Western | 37 |

***3.*** *imdb.sql database.*

Chosen because of the movie_ratings and movie basics tables within the sql database.

## DATA CLEANING

- tn.movie_budgets.csv.gz database.

This dataset was loaded into the **variable** ,*budget_df*. The Data-frame had no missing values, confirmed by the **.dtypes function.**

The main issues with this DataFrame was that some columns were in object data types which would not be useful for analysis.

I had to change the domestic_gross, worldwide_gross and production_budget to an int64 data type to enable numerical calculations.

The release_date was also converted from an **object data type** to a datetime64 data type.

- tmbd.movies.csv database.

This was loaded into the variable ,*tmdb_df*. The only issue this Data-Frame had was duplicated rows.

This was dealt with by defining a function ,get_duplicates() which removed any duplicates leaving only one version.

- *imdb.sql database.*

The database was an SQL file. I loaded the movie_basics table into a variable called movie_basics.

It had no duplicated values.

Missing values were observed , I decided to drop all rows with missing values

reducing the entries from **146144** to **112232** entries. Dropping was preferred as the missing values were vital for the analysis as we shall see in the data visualization sector.

Further Exploration of the data led to the discovery of an outlier. A movie with a runtime in minutes of 51420 , which is basically just short of 36 days.

I am as shocked as you are. The name of the movie is called Logistics , if you have the time, a Documentary just so you know.

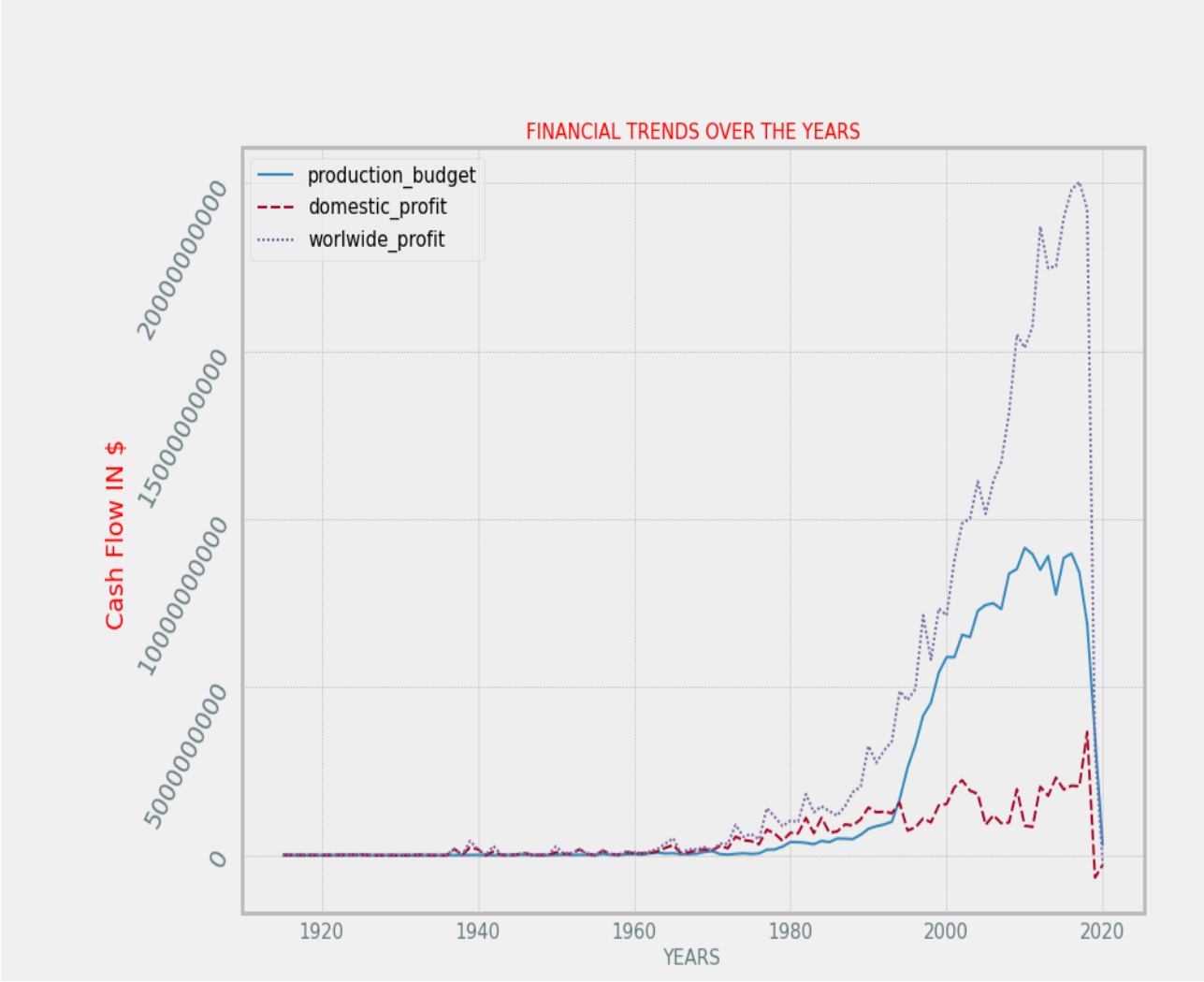
# *4. DATA ANALYSIS*

## *DATA VISUALIZATIONS*

For the graphs with languages and genres , the data available was a tremendous amount that they would have affected to readability of the graph hence I opted to limit them to the top 15 for languages and top 30 for genres.

### Focus on Foreign Audience.

*Figure 8: Trend of Finances over the years*



The line plot above follows the trend of cash flows in the movie industries from the late 1920's to 2020. I grouped the data by year of release for the movies and summed to make the Data Frame *"Budget_yearly".*

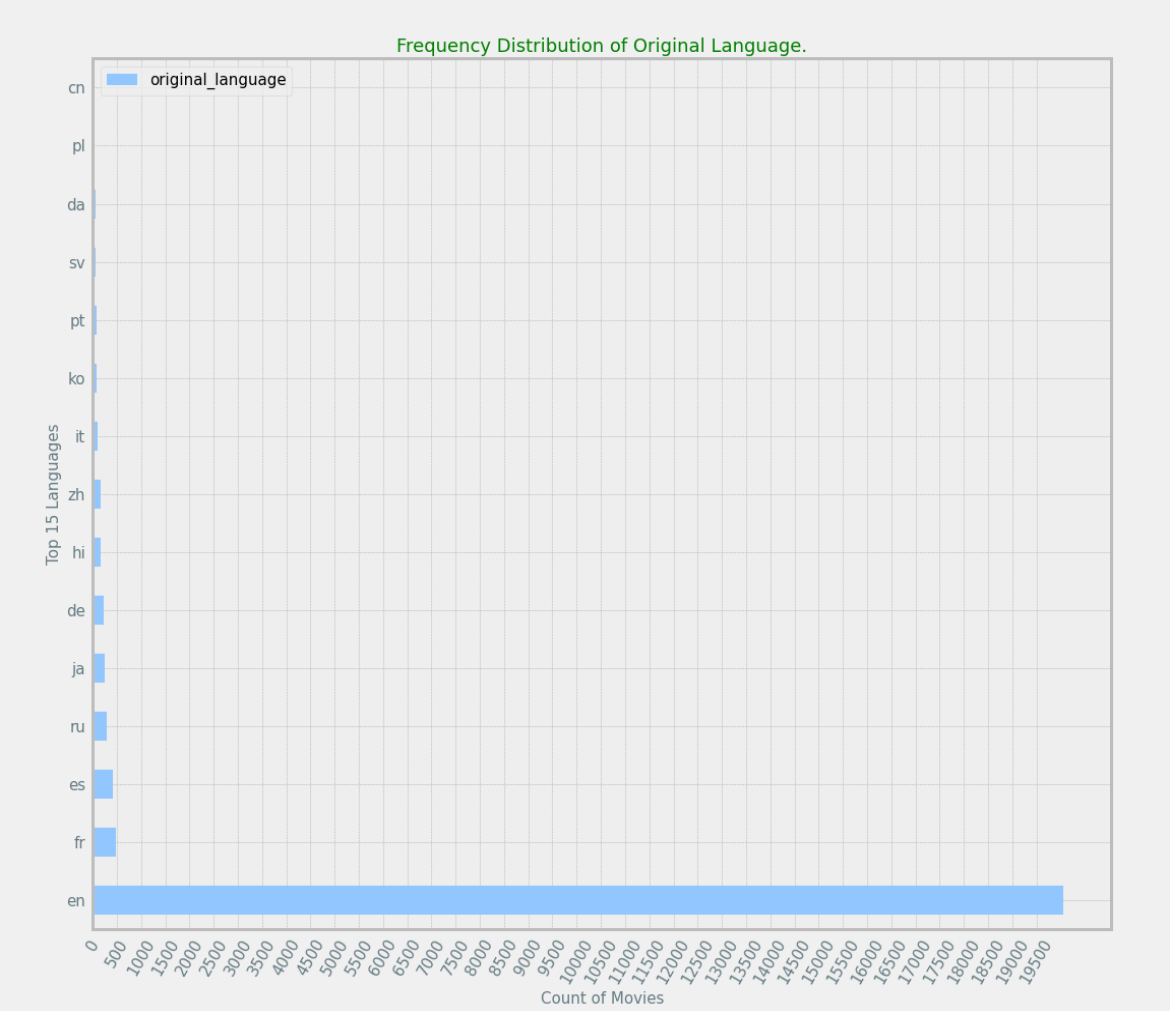
It is clear that the industry has been growing steadily and shot up between 1990 and 2000, especially the worldwide_profit, basically the international audience.

It should also be observed that the production budget is increasing with the profit, this is due to the improvement of technologies in the industry over the years.

And that the worldwide_profit surpassed the domestic profits. Completely.

The bar graph below is a representation of the top 15 movie counts based on languages.

*Figure 9: Frequency Distribution of the original Languages*




This shows that the most common language for movies is English (en), but the other languages should be considered they represent the international audience.

The top 14 languages include ;

- English (en)
- French (fr)
- Spanish (es)
- Russian (ru)
- Japanese (ja)

  German (de)
- Hindi (hi)

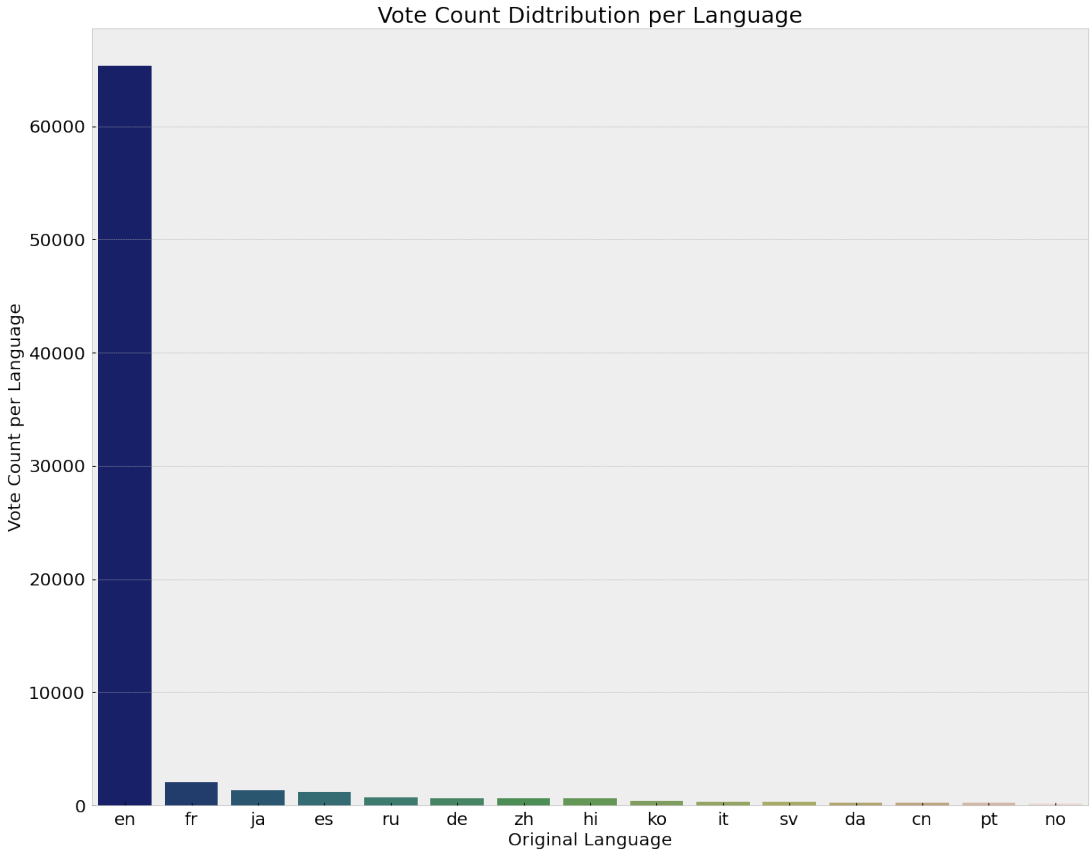
- Chinese (zh)
- Italian (it)
- Korean (ko)
- Portuguese (pt)
- Swedish (sv)
- Danish (da)
- Polish (pl)

Note the above is the distribution frequency of the languages.

More abbreviations from here.

A bar graph with the depiction of languages based on popularity follows.

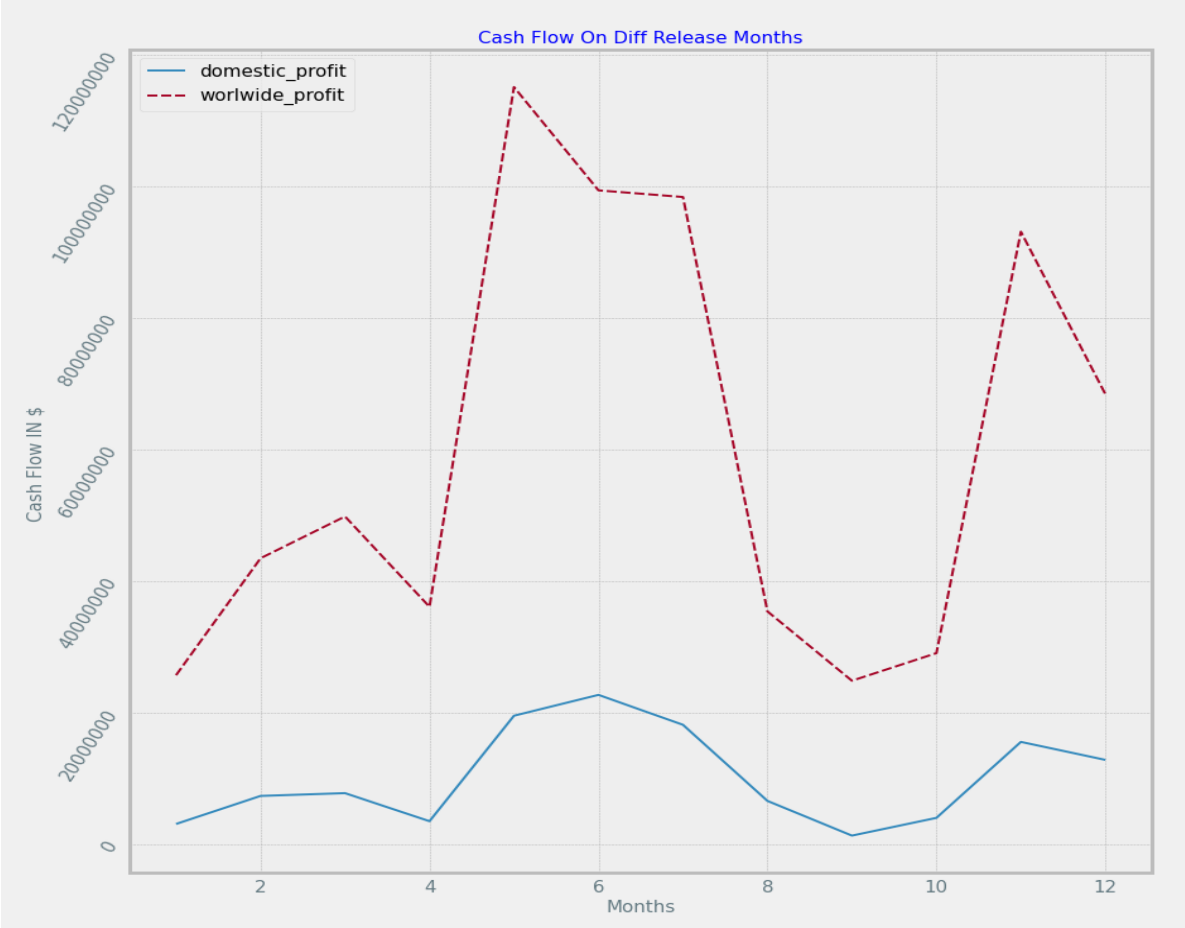
*Figure 10: Language Distribution based on Vote Counts*



It is clear in terms of vote counts there is not much difference from the two graphs. They are the same languages observed apart from the addition of Norwegian (no).

## Timing

*Figure 11: Mean Of Cash FLows over the years on the 12 months of the year*




There are consistent times of the month every year which yields lower profits when movies are released on, termed as 'Dump Months' and that there are preferred movie durations. These months are called dump months.

The Y-axis (Cash Flow IN $) is the mean of all the profits earned during the different release years.

The **dump months** are what the film community has, before the era of streaming-television.The two periods of the year when there have been lowered commercial and critical expectations for most new theatrical releases from American filmmakers and distributors. Domestic audiences during these periods are

smaller than the rest of the year, so no major movies are released. January and February are usually most commonly described this way. More on this here.

The graph above proves its existence, the most profitable season being between April (4th month) and August (8th month).

Another Data visualization on the aspect of Timing includes;

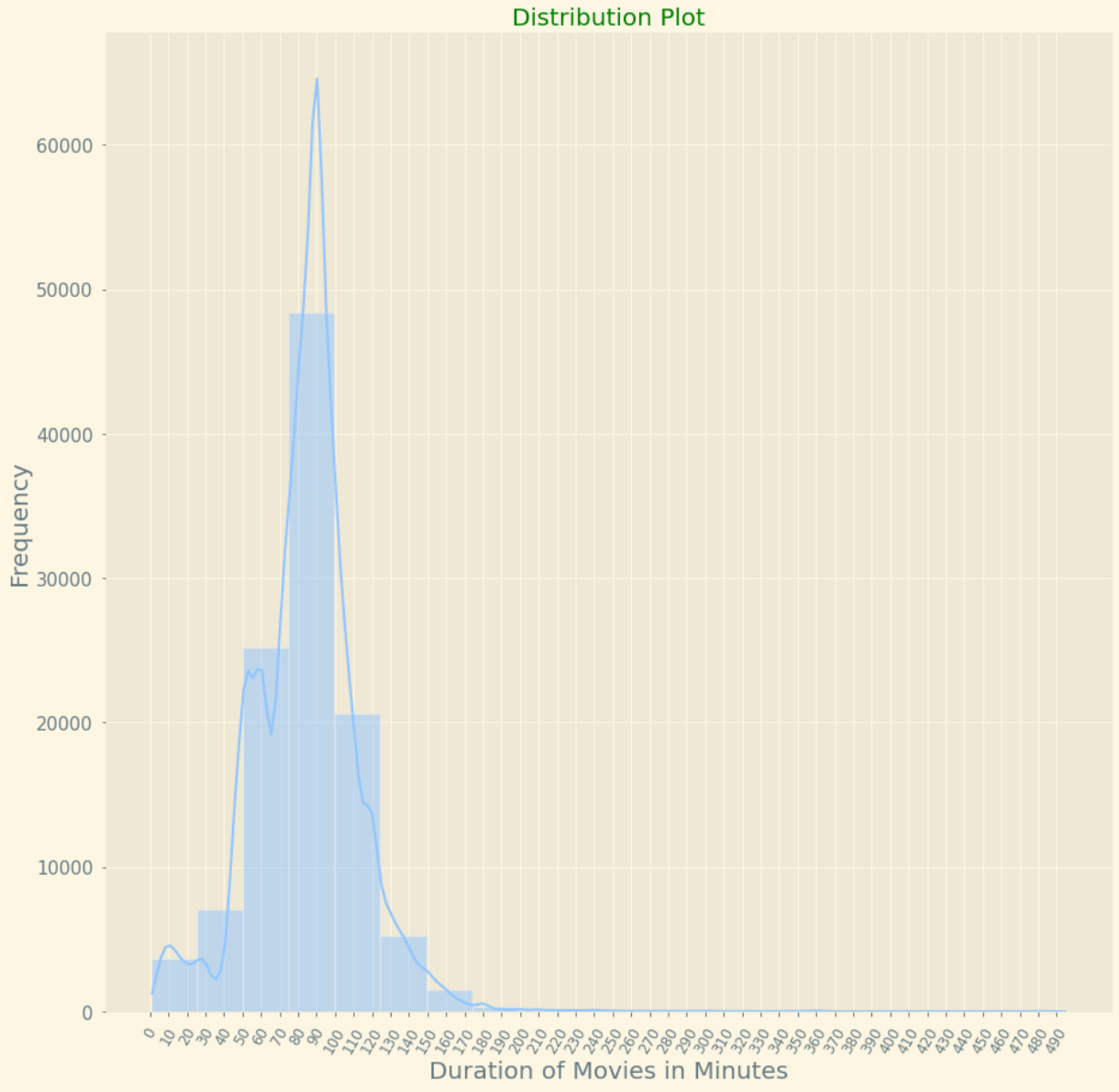
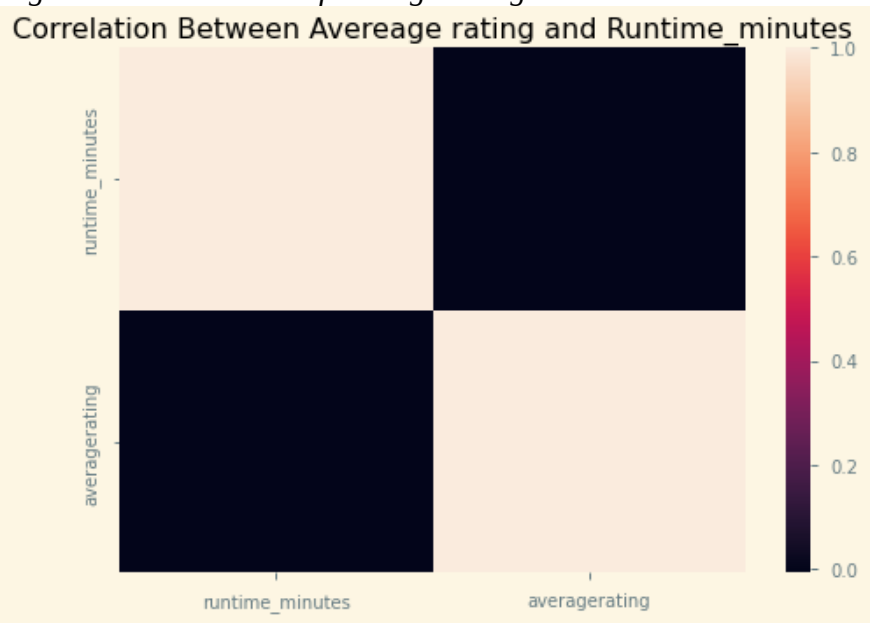



*Figure 12: Histogram of movie Runtimes in Minutes*

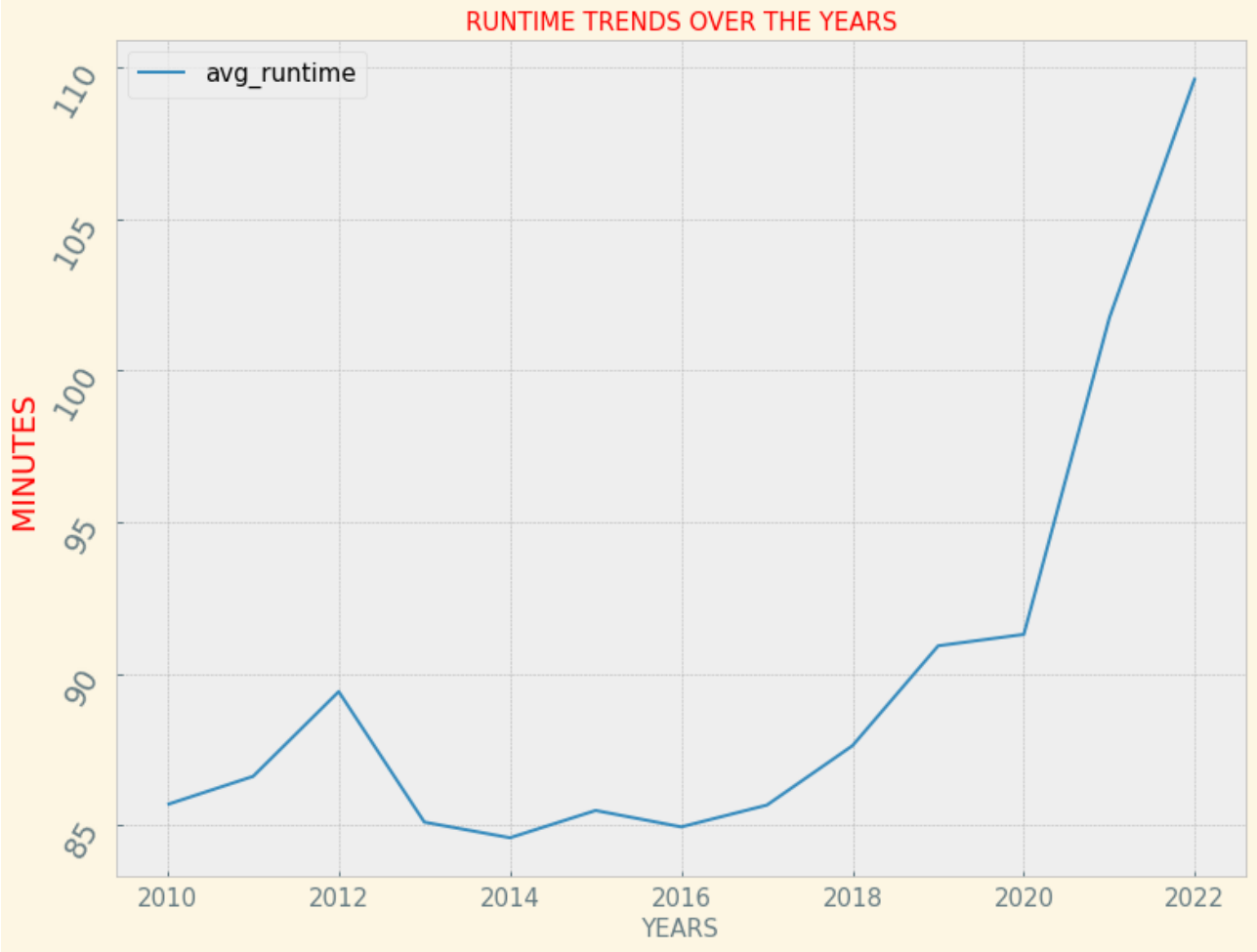
From the graph above we can tell it is a symmetrical histogram. A vast majority of movies are between 60 minutes to 100 minutes.

*Figure 13: Correlation of Average ratings and runtime minuets.*




The above graph is to show that the two have no positive correlation or negative correlation, meaning that the lengths of the films are purely subject to the set standards of the industry where a standard film is mostly 90 min long as shown from the graph. The exact corr between the two is [**-0.007075**].

*Figure 14: RUNTIME TRENDS OVER THE YEARS*

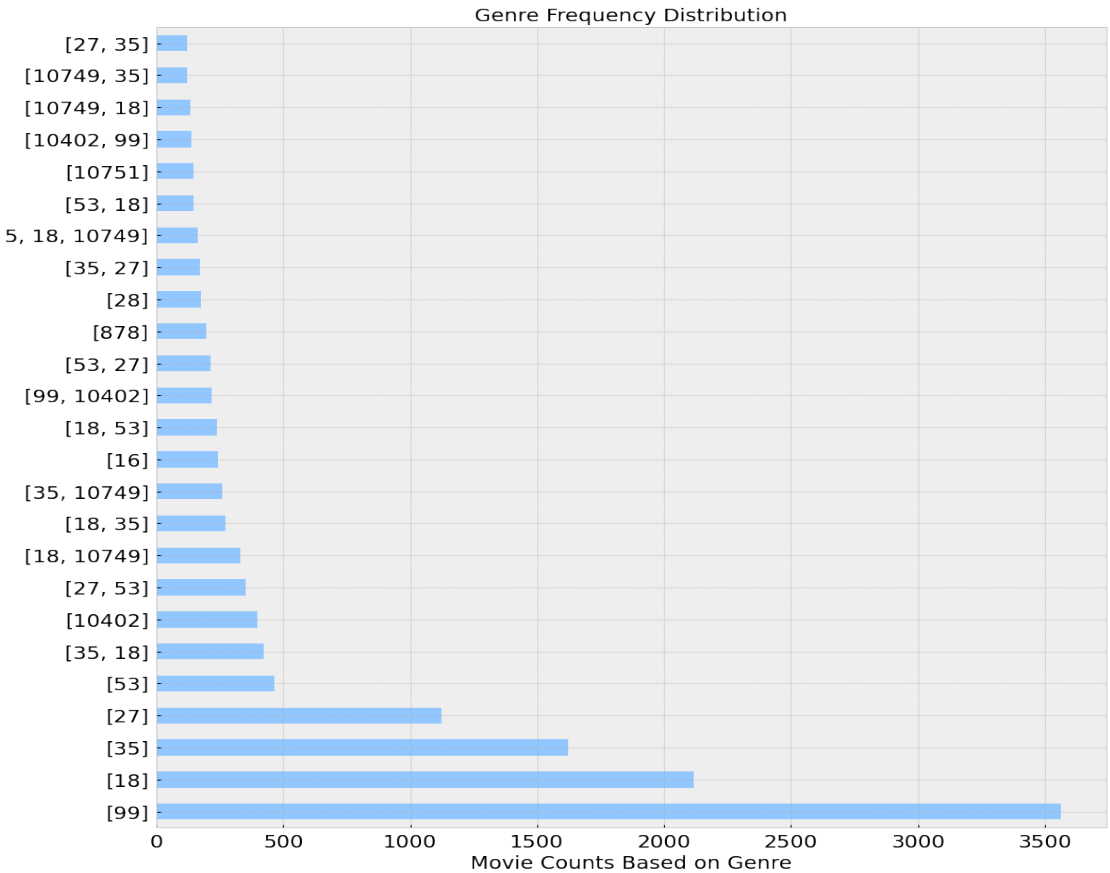



But it should also be noted from the above that the length of movies over the years have been on the rise over the past years.

## Genres

There are genre types which are preferred by the majority of movie watchers. The graph below Shows the bar-graph of the most popular genres based on there count from the movies in the tmdb_df Data-frame.

*Figure 15: Genre_id frequency distribution.*



From the graph above we can tell that the most common genre in the data frame is 99 which represents documentaries followed by Drama (18) and comedy (35).

The table with the references for the genre id is here  Genre key ID for the TMbd data.

The Graph that follows is similar to the one above, only difference is that they are from different databases.

But they still corroborate in terms of findings for the top 3 Documentary, Drama and Comedy, horror and Thrillers

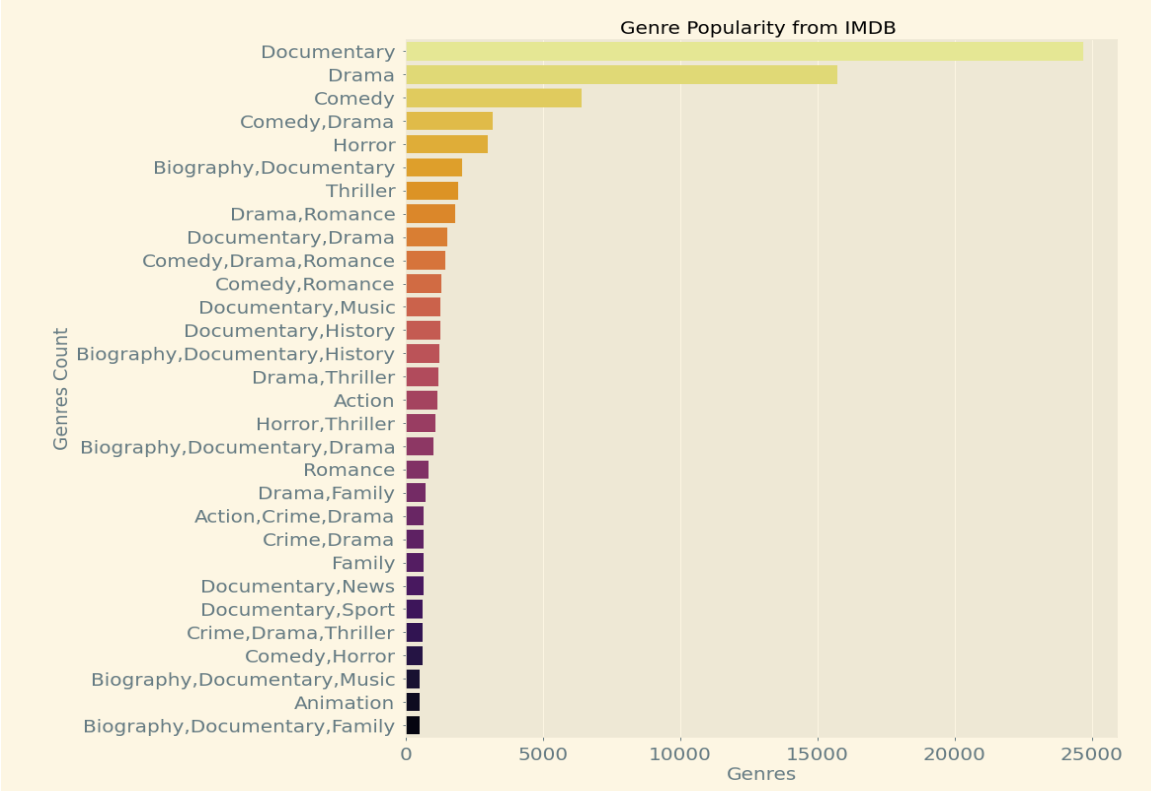
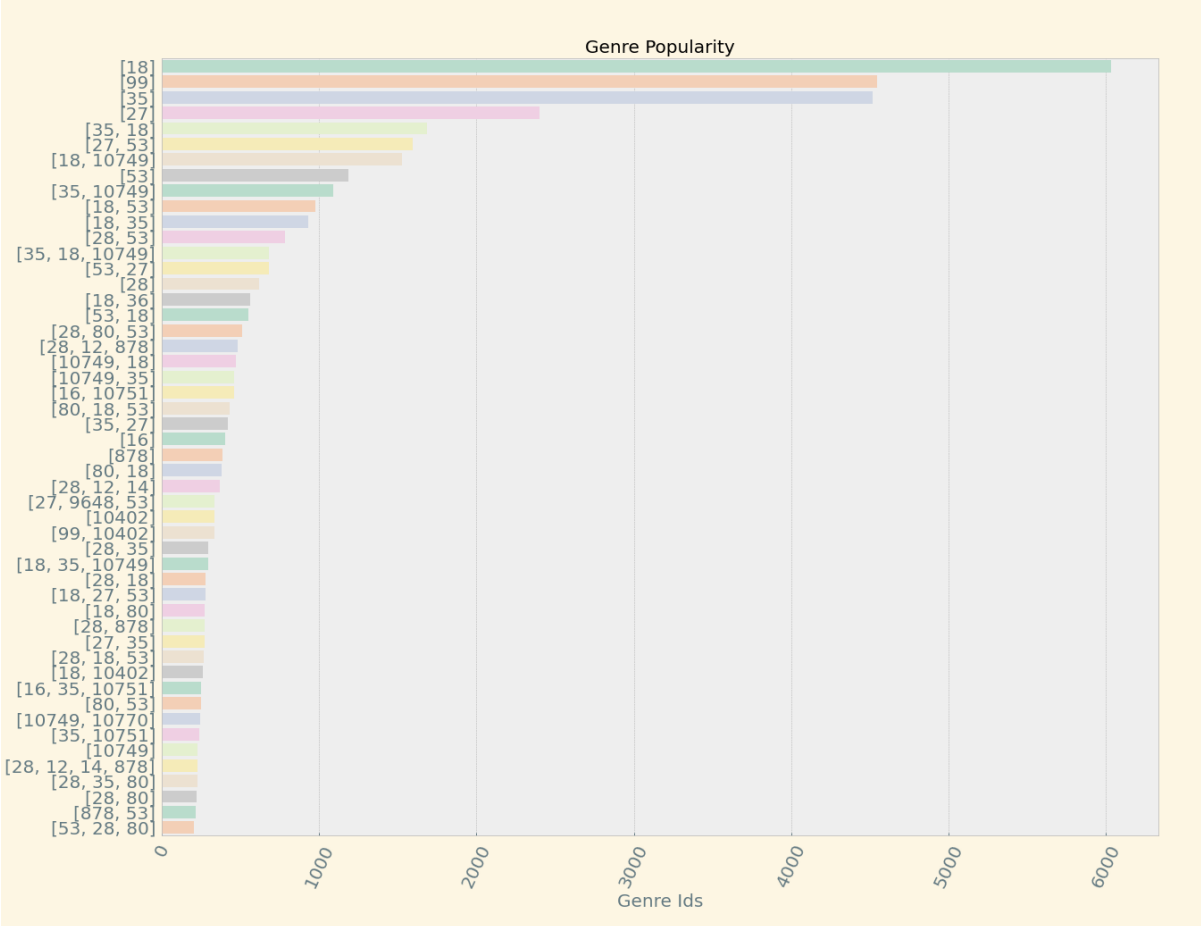


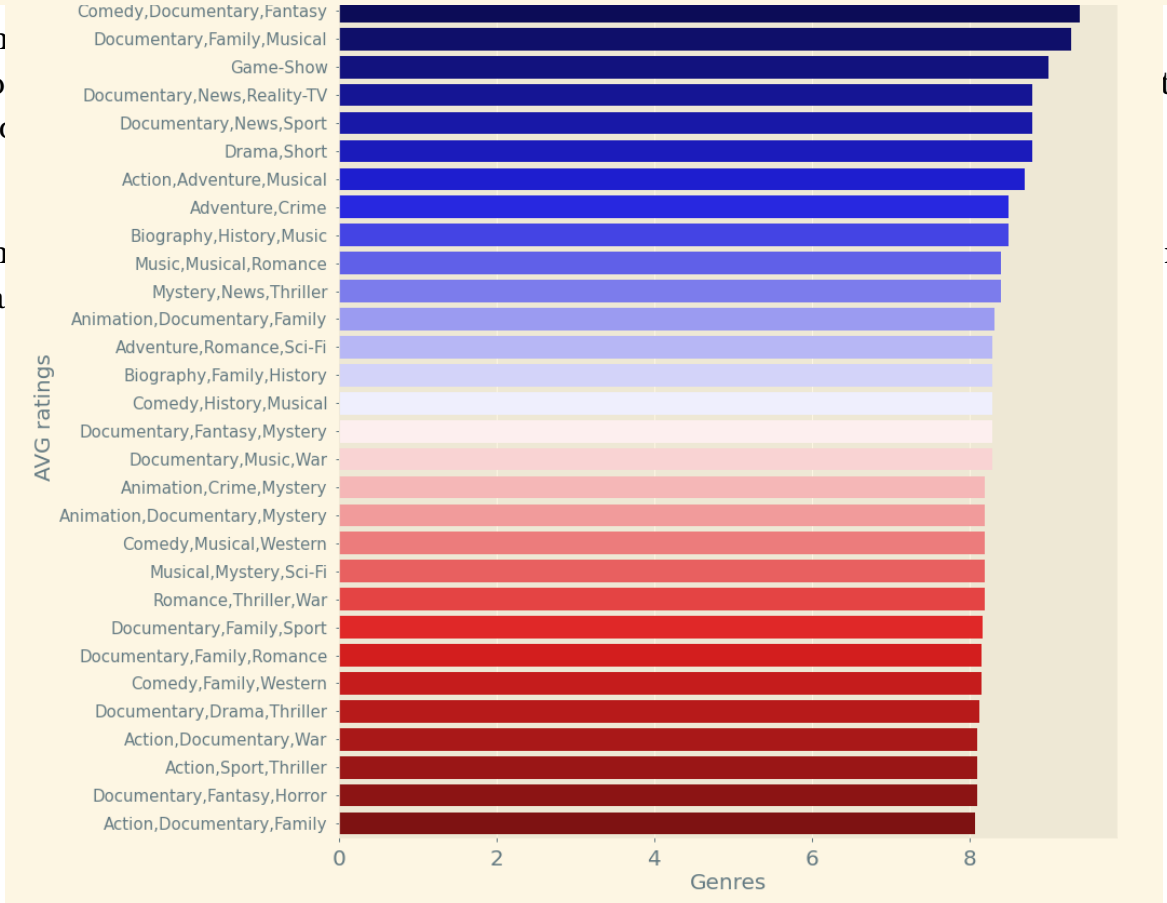

*Figure 16: Genre Frequency distribution for the top 15 genres.*

*Figure 17: Genre Id Bar-graph against their popularity.*




Th
po t
gr

Th imdb
da



# *5. CONCLUSION*

## HYPOTHESIS TESTING RESULTS.

### *The International market will be more profitable than the domestic and The English language is the most available language for movies.*

The International market has been more profitable than the domestic market this was shown by the line plot Trend of Finances over the years.

I have chosen to combined this with the language aspect, where most movies in the database were originally english, this was depicted in the Frequency Distribution of the original Languages and corroborated in the Language Distribution based on Vote Counts where English still dominated. the focus shouldn't be on the English but on the others that follow, as they are the international audience and therefore increase there number.

### *There are consistent times of the month every year which yields lower profits when movies are released on, termed as 'Dump Months'*

This was proven to be true in the graph Mean Of Cash FLows over the years on the 12 months of the year. The best months for profits is between April and August. And the worst months August to September when the schools open and January till March after the X-mass holidays. There existence varies from country to country as they are determined by the systems in place and cultural beliefs for holidays. More info here.

### *There is a preferred length of time for movies which the audience is used to, a set standard so to say.*

This was proven in the Histogram of movie Runtimes in Minutes. The most common times lied between 60 – 100 minutes but the peak following the density curve on the graph was 90 minutes.

A correlation graph was added to emphasize that the run time and the popularity of a movie were not correlated, the run time is entirely dependent on you as the Studio and the budget available.

Note also that there is a noted trend of increasing of the runtime in the past years.

***There are genre types which are preferred by the majority of movie watchers.***

This was proved using the two graphs , Figure 17: Genre Id Bar-graph against their popularity. and Figure 18: Genre Bar-gragh based on the Average of their ratings out of 10.(top 30).

Which both show that the audience have a prerogative towards certain genres with Dramas leading , from the tmdb database and a mixture of genres from the imdb database.

# *6. RECOMMENDATION*

I recommend that in the movies Microsoft will produce, Aside from english adding other language alternatives will go a long way in improving the movies odds for the international audience. This may be done either through Dubbing alternatives or Subtitles. A good example of a company that is doing this is Netflix but they take it a step further by getting movies from various countries and making it available to the rest of the world. As Sun Tzu wrote *"If you know the enemy and know yourself, you need not fear the result of a hundred battles."*

Microsoft should also avoid releasing there movies on the "Dump months" as they may go unnoticed by their target audience unless they are trying to release a movie which they don't know how it will be released the dump months can be used like so.

The movie genre that is really popular from the tmdb is Drama's , comedy ,Documentary, Horror and thriller. Those would be a good start, the imdb genre by popularity gives an example of the top, From those two list a good understanding of what is loved can be understood.

It should be noted that this database was based on websites from the US which means a lot of the graphs are biased towards that country, I recommend a future research similar to this but on other countries movies data base to get a good view of the international audience.
Another point of interest is that this database is 2 years old and an up to date research to be done so us to increase the validity of its results.