**Introduction**

Movies!! When we think movies the first words that may come to mind are action, romance, comedy, and horror! However, here we take a cold, hard, statistical look at the magic called movies. We have sliced and diced the data and crunched few numbers. It is for you to decide if statistics comes out as the ever powerful super-hero who can do no wrong or a flawed hero, but hero nonetheless.

We have used the popular and reliable website <http://www.omdbapi.com/>

to source our data. We used IMDB’s APIs and obtained data on movies from 1909 to 2019.

A commercially successful movie not only entertains audiences, but also enables film companies to gain tremendous profit. A lot of factors such as excellent directors, creative storytelling, a star-studded cast, and large budget can contribute to the making of a blockbuster film.

Movie ratings and reviews at sites such as IMDb are commonly used by moviegoers to decide which movie to watch, rent, or buy.

The objective of this exercise is to take a cold hard statistical look at the business of movie making. It will also challenge or ascertain popular beliefs on popularity, salability.

# **Objective**

This project analyzes and interprets data on the IMDB website for movies released between years of 1910 to 2019. The following three questions were examined during the course of the analysis:

1. How has movie revenue and popularity changed over time?
2. Where are movies produced? What languages are used in the movies?
3. Can data be displayed non-numerically? When and why?

the movie data on the OMDb website over 100 years (released in 1910-2018) with graphics and gives an interpretation of these data.

**Preparing, Exploring, and Cleaning up the Dataset**

We started by getting the names of over 4000 movie titles from the TMDB 5000 Movie Dataset from Kaggle. Next, we extracted information related to these movie titles using the API from the OMDB website (<url:https://www.omdbapi.com>).

The Open Movie Database API allows you to search movies by title and returns JSON items for Title, Release Year, IMDB ID, Type (movie, series, or episode), and Movie Poster image. Within each of these JSON items, we extracted values from the following columns to arrive at the final data that we wanted to analyze and saved it into a csv file:

1. Title – Movie Title
2. Year – Movie released year
3. Rated – MPAA viewer rating
4. Runtime- Duration of Movie in minutes
5. Genre – movie genre
6. Director – director (s) of movie
7. Writer – movie writer (s)
8. Language – spoken languages in movie
9. Country – country/countries of production
10. Awards – awards received
11. IMDB Rating (score out of 10)
12. Revenue – movie revenue \*
13. Budget- movie budget \*

(Note: Items 12 and 13 were extracted from TMDB Kaggle dataset which was merged with the data set created with API calls.)

Some of the steps we took to clean the final CSV file included:

1. Dropping columns with a significant amount of missing data.
2. Reviewing columns for inconsistent data and either dropping or converting the data to the desired data type.
3. Filling missing data with either NA or 0 depending on the type.
4. Exporting a final cleaned .csv file.

For a more detailed explanation of the clean-up process, please see the project Jupyter Notebook.

**Data Parsing:**

There were also certain rows that contained lists of information, such as the “Genre”, “Language Columns,” “Country,” and “Awards.” The items in these lists needed to be parsed and stored in a larger list so that their values could be counted and otherwise manipulated.

# **Analysis**

# **How has movie revenue and popularity changed over time?**

Movie revenue has increased over a period of time. Movie revenue depends on lots of factors like budget, genre, story, popularity, actors, production company, actors, directors etc.

Movie revenue comes from factors such as ticket price revenue, merchandising Dollars (It all started with Star Wars), foreign sales, television rights, VOD video-on-demand (VOD) and streaming.

Here are some observations:

Presenting average income per decade. Since this as raw dollar figures, not adjusted for inflation we can see the expected trend of the numbers rising per decade.

There are 3 interesting trends   
   a. Revenues in decade of 1930 remains unmatched till 1970  
   b. Percentage rise in current decade far exceeds the rise in last couple of decades  
   c. Similar rise can be observed in decades of 70s and 80s

Movies that are well rated by public and critics make the most money, There is a significant raise in budget over period of time.

There does not appear to be a relationship between budget and runtime of movies. Movies with a budget less than $10 million are more than 2 hours long.

The revenue vs. runtime regression analysis also shows no apparent relationship between revenue and runtime. But there are more outliers than in budget vs. runtime. Movies running more than 120 mins generate revenue <$10 million, indicating that the general public do not enjoy longer runtimes.

A negative t test and a p value < 0.05 for budget and revenue data show a significant difference between budget and revenue data. The mean revenue ($84,385,511.88) is greater than the mean budget ($29,620,517.88), indicating an overall profit for most movies

**Where are movies produced? What languages are used in the movies?**

In the dataset, the “Country” field represents the countries where the production companies who financed the film were located. This means, for example, even if a title is shot on location in France, if its production companies are all based in the USA, the country is recorded as USA. The “Languages” field records each language that is spoken in each title. Thus, for example, if a movie like “Bend it Like Beckham,” has languages like Hindi, English, German, Punjabi, and English listed, it means that there are some words spoken in each of those languages in that film.

In 2019, the movie Roma won Oscar awards for “Best Foreign Language Film” and “Best Director.” Roma was a film set in Mexico about a housekeeper in a middle-class family in the 1970s. In his speech for “Best Foreign Language Film,” the director, Alfonso Cuaron, stated:

‘I grew up watching foreign language films and learning so much from them and being inspired. Films like ‘Citizen Kane,’ ‘Jaws,’ ‘Rashomon,’ ‘The Godfather’ and ‘Breathless.’

The director’s speech asked us to consider why the label of “foreign” is attached to some movies but not others and why all films irrespective of language are not considered the same.

The breakdown of the languages and countries in the dataset can be used to get a better understanding of how and why a category of “best foreign language film” might come into existence.

The data showed that in our dataset of over 4000 movies, English was used in 93.98% of the movies, or in over 4000 of the movies. The next most frequent language that appeared was Spanish at 9.58 % (a little more than 400 movies), and French at 8.51% (just under 400 movies). The frequency of English is almost nine times more than the frequency of Spanish. Also, another interesting fact was that even though China is the world’s most populous country, in the dataset that we had, Mandarin and Cantonese phrases were only spoken in under 2% of the movies which were represented.

In terms of the number of languages used in each movie, the data showed that in 66.6% of the movies we had, only one language was spoken. There were two languages spoken in 20.4% of the movies, and three languages spoken in 8.4% of the movies. This suggests that the number of different cultures being represented in movies is low.

The country analysis turned out similar to that of languages in that in the dataset there are 4020 movies that were produced in the United States, with the United Kingdom coming in second at 738, Germany at 360, and France at 340.

The disparity in numbers suggests that on the one hand, when there are so many movies of one type, not having a separate category for movies gives them less of a chance to win and be showcased, and therefore draw others to create movies of a similar type. On the other hand, there often risks associated with categorizing as well.

**Can data be visualized without using numbers? When and why?"**

Word cloud insert first shot - "Yes, data can be visualized without using numbers with a visualization called a word cloud. A word cloud allows you visualize the frequency with which an item appears in a set of data or show the importance of an item in a set of data by using the font size or color of the word to indicate its weight or significance. In our analysis, we used word clouds to show the frequency with which a language or country was associated with the movies in our dataset, and also (Poonam - describe other word clouds used). (Kit - Insert anything technical about generating them if you want). Word clouds are great tools to use in posters, on a website, or in presentations where the audience you are presenting to does not have a long time to study your visualization."

# **Conclusion /Takeaways**

# This study through a large volume of data, determine the following points for movies released between 2010 and 2018:

* Genre drama is the most frequent and revenue earning genre.
* Percentage rise in current decade far exceeds the rise in last couple of decades
* Number of movies produced and profit increased with each passing decade(without adjusting for inflation), we should account for inflation to get more realistic analysis.
* Most movies last between 60 minutes and 120 minutes
* Movies that are well rated by public and critics make the most money
* The more the public appreciates a film, the more they vote and give a good rating
* There is a significant raise in budget over period of time.
* Average IMDB rating has gone down over period of time.
* Action, Comedy and Horror movies have the most rated R movies; the directors of these films appear to be catering to a more mature audience. Rated PG-13 comes in second highest.
* There does not appear to be a relationship between budget and runtime of movies. Movies with a budget less than $10 million are more than 2 hours long
* The revenue vs. runtime regression analysis also shows no apparent relationship between revenue and runtime. But there are more outliers than in budget vs. runtime. Movies running more than 120 mins generate revenue <$10 million, indicating that the general public do not enjoy longer runtimes.
* A negative t test and a p value < 0.05 for budget and revenue data show a significant difference between budget and revenue data. The mean revenue ($84,385,511.88) is greater than the mean budget ($29,620,517.88), indicating an overall profit for most movies