# **Introduction**

Movies!! When we think movies the first words that come to our minds are action, romance, thrill and what you may. However here we want to take a cold hard statistical look at the magic called movies. We have sliced & diced the numbers, crunched few numbers. It is for you to judge if statistics comes out as ever powerful super-hero who can do no wrong or a flawed hero, hero none the less.

We have used popular and reliable website [www.imdb.com](http://www.imdb.com) (Internet Movie Database) to source our data. Data is sourced using their APIs and we were able to fetch data on movies since 1910. Movies are split across genres, language, ratings , the duration of the film and many other factors including the obvious ones like actors, director, year, revenue etc.

A commercial successful movie not only entertains audience, but also enables film companies to gain tremendous profit. A lot of factors such as good directors, story, cast along with a good budget are considerable for creating good movies.

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

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 aims to analyze the movies data on the IMDb website over 100 years (released in 1910-2018) with graphics and gives an interpretation of these data.

# **Preparing the dataset**

Data is extracted using API from IMDb website, here is the URL

<https://www.imdb.com>

Here are the data we are interested in:

* Title
* Year
* Rated
* Released
* Runtime
* Genre
* Director
* Writer
* Language
* Country
* Awards
* Ratings (score out of 10)

**Data Cleaning**

This dataset can be fixed as follows:

1. Delete the line with the missing values
2. Fill empty fields with specific values
3. Fill empty fields with calculations

We went with option 1 & 2

# **Data analysis**

# **Visualization of data and interpretation of data**

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

* Drama is the most frequent and revenue earning genre.
* Average revenue of movies increased exponentially with each decade.
* Number of movies produced and profit increased with each passing decade.
* 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
* Movies
* Movies that exceed 3 hours bring in the least money
* 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