## AIM: To create a model that analyzes the ratings on IMdb and further draws correlation through the understanding of this same data

## Introduction:

## Hundreds of thousands of movies have been released since the ship set sail with the first ever movie in the late 1800s. Predicting the success of a movie is a complex task as various factors influence its performance in the box office. Since a huge capital is invested in the production, marketing, promotion and distribution of the movies, it has been a topic of interest not only for the viewers, but also for the media and production houses and all others who are involved in these processes. For the study, a part of the IMDB dataset will be used and scrutinized to get valuable information. In this digital age, online publicity plays a major role in the success of a movie. This is hence a field that would benefit from a study that helps them predict/ analyze the likelihood of the success of a movie based on the following attributes.

IMDB rates movies according to true Bayesian estimate. WR = (v ÷ (v+m)) × R + (m ÷ (v+m)) × C (1) where:

R = Mean Rating

v = votes for the movie

m = minimum votes required to be listed in the Top 250 (currently 25000) C = the mean vote across the whole report (currently 7.0)

OBJECTIVE AND SCOPE:

o Identify the sample size

o EDA (exploratory data analysis)

o Cleaning data

o Perform PCA

o K means

o DBSCAN

o Random Forest Technique