FUTURE VISION™
COMPUTER INSTITUTE

# PROBLEM STATEMENT

IMDB Movie dataset Data Analysis

# PROBLEM STATEMENT


The problem at hand involves analyzing a dataset of IMDB movies to extract valuable insights into the performance and characteristics of movies across various dimensions such as runtime, ratings, revenue, and voting patterns. The dataset includes key attributes like movie title, genre, director, actors, year of release, runtime, rating, votes, revenue, and metascore, which provide a rich foundation for addressing multiple business questions.

One key problem is identifying movies with exceptionally long runtimes (greater than or equal to 180 minutes) and understanding whether these movies perform differently in terms of ratings, revenue, or audience reception. Additionally, the dataset can help identify trends in voting activity, such as which years saw the highest number of votes, offering insight into periods of heightened audience engagement.

Another crucial aspect of the analysis is examining movie revenue over time, which can reveal which years were more financially successful, and whether those years correlate with particular blockbuster releases. A further analysis of director performance, by calculating the average ratings of movies by each director, could help assess the consistency and success of various filmmakers.

Understanding the number of movies produced each year could also provide insights into industry trends, such as whether the volume of movie production is increasing or decreasing and how that might affect overall market dynamics.

# BUSINESS PROBLEM OVERVIEW

The business problem presented by the IMDB movie dataset revolves around extracting actionable insights from a variety of movie performance metrics, with a focus on understanding the relationship between key factors such as movie runtime, ratings, revenue, and audience engagement.

The primary challenge lies in analyzing and leveraging these diverse attributes to inform strategic decision-making for studios, production companies, and distributors. By examining the dataset, businesses can uncover trends such as which types of movies—based on runtime, genre, or director—perform better in terms of revenue and audience ratings.

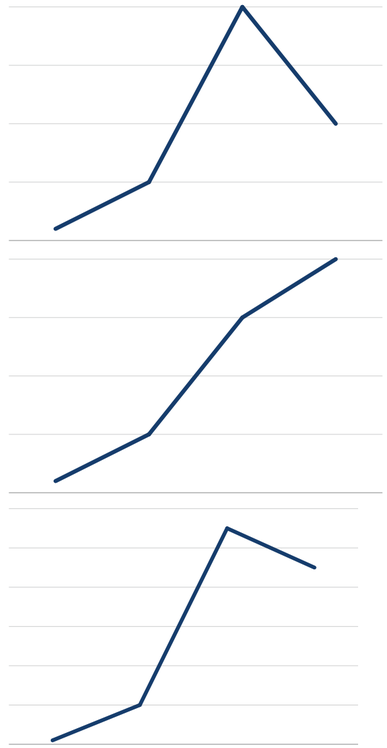
For example, identifying movies with runtimes of 180 minutes or more could offer insights into whether longer movies correlate with higher box office returns or greater audience engagement.
The business problem also extends to understanding the relationship between movie ratings and financial performance. By analyzing whether higher-rated movies tend to generate higher revenue, companies can better tailor their marketing efforts, focus on producing higher-quality films, and refine their audience targeting strategies.

The classification of movies into categories such as "Excellent," "Good," and "Average" based on ratings allows for more precise segmentation and can inform content recommendations, helping businesses cater to specific audience preferences.

The IMDB movie dataset refers to a collection of detailed information about various movies, including attributes such as title, genre, description, director, actors, year of release, runtime, rating, number of votes, revenue (in millions), and metascore. This dataset is used to analyze movie performance and identify trends in factors that influence a movie's success, both critically and commercially.

Each movie in the dataset is represented by a set of attributes, where the title and genre describe the content of the film, while the runtime and rating provide insight into its duration and critical reception.

The inclusion of revenue and votes allows for an evaluation of how the movie performed at the box office and how audiences rated it on the IMDB platform. Directors, actors, and metascores give a deeper view into the creative and critical aspects of the film, while the year of release enables temporal analysis to identify trends over time.



# UNDERSTANDING & DEFINING DATASET

# PROJECT PIPELINE

The project pipeline can be briefly summarized in the following steps:

- Data Understanding: Here, we need to load the data and understand the features present in it. This would help us choose the features that we will need for your final model.

- Exploratory data analytics (EDA): Normally, in this step, we need to perform univariate and bivariate analyses of the data, followed by feature transformations, if necessary. For the current data set, because Gaussian variables are used, we do not need to perform Z-scaling. However, you can check if there is any skewness in the data and try to mitigate it, as it might cause problems during the model-building phase.

FUTURE VISION™
COMPUTER INSTITUTE

# THANK YOU