# **IMDb movies extensive dataset**

A picture containing text, bunch, posing, various

Description automatically generated

**Content**

IMDB (Internet Movie Databases) is the most popular movie website. This website is well known for storing almost every movie that has ever been released.

IMDb stores information related to more than 6 million titles, and it is owned by Amazon since 1998.

I am personally interested in this dataset as I am a movie fan.

This dataset includes 85,855 movies with attributes such as movie description, average rating, number of votes, genre, etc.

This dataset can be found at Kaggle in this link: [IMDb movies extensive dataset | Kaggle](https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset). Data has been scraped from the publicly available website [https://www.imdb.com](https://www.imdb.com/).

The dataset is available as the .csv file. a sample of data is shown in the following table:

**A picture containing text

Description automatically generated**

**Data description:**

|  |  |
| --- | --- |
| Column name | Column description |
| imdb\_title\_id | title ID on IMDb |
| title | title name |
| original\_title | original title name |
| year | year of release |
| date\_published | date of release |
| genre | movie genre |
| duration | duration (in minutes) |
| country | movie country |
| language | movie language |
| director | director name |
| writer | writer name |
| production\_company | production company |
| actors | actor names |
| description | plot descrption |
| avg\_vote | average vote |
| votes | number of votes received |
| budget | budget |

**Objective:**

* Analyse the data available on the IMDb website for movies.
* I am looking forward to exploring various visualization techniques like charts, graphs and word clouds.

**Questions:**

From this dataset I am curious to find out:

* What is the top (15) countries producing movies content?
* What is the top (15) Languages in which movies are produced?
* What is the genre of the most movies made in?
* Visualize the rise in movie content released each year and the drop in content because of COVID-19 Pandemic.
* What is the shortest movie in IMBb?
* What is the longest movie in IMBb?
* What is the list of short movies (shorter than 60 minutes)?

**Model:**

1. Predict the number of movies will produce next year.
2. Predict the average vote the new movie will have depend on some of these features (genre, production company, director, writer, language, country).

**Tools**

There are tools that will be used to achieve the goal of this study, such as:

***numpy, pandas, matplotlib and seaborn*** for discovering and showing the data and train a model. The work will be done through ***Jupyter notebook*** using ***python***.

**TO DO:**

* Explore the data and come up with EDA phases then use a model to fit the data.
* **NOTE:** the used features may be increased or changed and the model as well.