# **Analysis Movies Industry**

A picture containing text, bunch, posing, various

Description automatically generated

*By: Mona AL-Nasser*

**Objectives:**

The goal of this project is to

1-Exploring the movies industry and finding answers for some questions such as:

* What is the top (10) countries producing movies?
* What is the top (10) Languages in which movies are produced?
* What are the most genres are movies made in?
* Which year was the best year of production movies?
* What is the list of short movies (shorter than 60 minutes)?
* What is the shortest movie in IMDb?
* What is the longest movie in IMDb?
* Extract the movies produced each year.

2- Create a model to predict the number of movies will produce next year.

### Design:

I worked on IMDB (Internet Movie Databases). It is the most popular movie website. This website is well known for storing almost every movie has been released.

First, I download the data from Kaggle website and read it. Then, I cleaned and explored the data. After that I showed the data to find the relation between the features. Then, implemented a regression model to predict my target, and split the data to fit and test the model.

### Data:

The dataset contains 85,855 movies with 22 attributes such as:

*movie title, year, genre, duration, country, average rating, etc.*

This dataset can be found at Kaggle in this link: [IMDb movies extensive dataset | Kaggle](https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset).

It’s available as the .csv file. a sample of data is shown in the following table:

**A picture containing text

Description automatically generated**

### Algorithms:

I started to clean the dataset by:

* Keep only the columns I need it (I kept only 12 columns)

(Title, year, genre, duration, country, language, director, writer,

Production company, actors, average vote, votes)

* Dropped all the missing values.
* Dropped the duplicated rows.
* Deleted any movie released before year 2000.
* Changed the type of column(year) to integer.

Then, I did some exploration on the data to find the answers for the question above. After that, I implemented a model to predict the number of movies will produce next year. I used a ***Regression\ Supervised\ Machine Learning algorithms*** such as:

(**LinearRegression, Ridge, and PolynomialRegression)**

These models take the year and predict the number of movies produced in this year.

To see the performance of these models I split my data into 80%(train- validation) sets /20% test set, and fit the models on train set, and test it on the validation and test sets.

The best Model to predict the target was the ***simple linear regression*** with simple train and test techniques it’s score R^2 = 0.719 on the test sets.

This are the scores of all the models I did:

* the simple linear regression with simple train and test techniques 🡪

R^2 = 0.719 on the test

* linear regression with (train\Validation\test) 🡪 R^2 = 0.29
* Ridge Regression on test 🡪 R^2: 0.40
* Cross-Validation 🡪 0.6856
* Degree 2 polynomial regression on test🡪 R^2: 0.664

Note:

I intend to build another model for predict the average rating for the new movie depend on some movie’s features such as:  
 *movie’s genre  
 production company  
 movie’s language*

I started to explore the relation between the average rating and (Movie’s genre and the production company). I plotted the charts and encoded the categories value to apply the suitable model. I tried Linear, Polynomial, and random forest regression, but unfortunately, I didn’t get a good scores.

### Tools:

There were many tools that I used to achieve the goal of this study, such as:

* ***numpy, pandas*** for discovering and cleaning the data.
* ***matplotlib and seaborn*** for showing the data.
* ***sklearn*** for building and training the model.

The worked done through ***Jupyter notebook*** using ***python***.

### Communication:

### The chart below shows the relation between the year and number of movies produced each year:

Chart, scatter chart

Description automatically generated

After applying the linear regression model. We can see in the chart below the actual number of movies, and the number of movies the model predicts for each year.

Chart, histogram

Description automatically generated Table

Description automatically generated