# IMDB Movie Analysis

**Final Project-1**

Difficulty Level: 5/5

**Description:**

For your Final Project, we are providing you with dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.

We can do this by asking 'What?' This is where you frame the problem i.e. What is the problem?

Use these questions to guide your thinking:

- What do you see happening?
- What is your hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Answering these questions will help you define a problem you are trying to solve and will allow you to find the right data to solve it.
Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

Make sure to use 5 Whys Analysis in your analysis and use this to create a report which conveys a data story.

Once you have framed the problem and gathered initial insights from the data, you can ask the following questions as you dig deeper into your analysis.

- What do you see happening?
- What are the specific symptoms of the problem?
- What is your hypothesis for the cause of the problem?

Five 'Whys' approach

Once you have the problem better defined, you can use 5 Whys technique to determine its root cause by repeatedly asking the question “Why”.

It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries. Here’s an example of how this technique could be used to figure out

the cause of the following problem: A business went over budget on a recent project.

Q: “Why did we go over budget on our project?”

A: It took much longer than we expected to complete.

Q: “Why did it take longer than expected to complete?”

A: We had to redesign several elements of the product.

Q: “Why did we have to redesign elements of the product?”

A: Features of the product were confusing to use.

Q: “Why were the features of the product confusing to use?”

A: We made incorrect assumptions about what users wanted.

Q: “Why did we make incorrect assumptions about what users wanted?”

A: Our user experience research team didn’t ask effective questions.

As you see above, what looked like a budgeting problem turned out to be a problem with the user experience team not working effectively.

While asking Why is easy, what we're interested in is the answer. Each time you answer why the next time gets more difficult as you must think deeper behind the reasons for this. As you ask why, you may find that you have multiple answers for the same question.

You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

A. **Cleaning the data::** PThis is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
   **Your task:** Clean the data
B. **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
   **Your task:** Find the movies with the highest profit?
C. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also

make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
**Your task:** Find IMDB Top 250

D. **Best Directors:** TGroup the column using the director_name column.

Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
**Your task:** Find the best directors

E. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
**Your task:** Find popular genres

F. **Charts:** Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

**Your task:** Find the critic-favorite and audience-favorite actors