PROBLEM STATEMENT

Welcome to week 3 of the Internship Program (IP) 2021-22. I hope you guys are enjoying the experience of the IP 2021-22 Internship so far. In this week, you need to work on building a mini project for Data Science.

To solve any Data Science problem the important asset is a good dataset. But we are living in a world where you don't get anything so easily so how could you expect to get a dataset clean and ready to use to solve the problem. The dataset comes with missing values, duplicate values, incorrect, incomplete, irrelevant, duplicated, or improperly formatted data. You need to clean the datasets and perform the actions mentioned below.

Here, we ask you to perform the analysis using the Exploratory Data Analysis technique. You need to find features affecting the ratings of any particular movie.

Analysis Tasks to be performed:

Explore the datasets using visual representations (graphs or tables), also include your comments on the following:

1. User Age Distribution
2. User rating of the movie "Toy Story"
3. Top 25 movies by viewership rating
4. Find the ratings for all the movies reviewed by for a particular user of user-id = 2696
5. There should the different graphs used for visualizing the dataset.

Use column genres:

1. Find out all the unique genres (Hint: split the data in column genre making a list and then process the data to find out only the unique categories of genres)
2. Create a separate column for each genre category with a one-hot encoding (1 and 0) whether or not the movie belongs to that genre.
3. Determine the features affecting the ratings of any particular movie.

Dataset:

https://cloudcounselage0-my.sharepoint.com/:u:/g/personal/welcome_cloudcounselage_com/ERaO65HV1elBiCLlH9gZlacBI30j1RohD7YLGWqWvtQRYA?e=68umZc

Note:

1. You are free to use any programming language.
2. The project will be judged based on the ability of the algorithm to perform data cleaning of the dataset.