**PROJECT: MOVIE-LENSE**

*Submitted by: Mr. Tejas Phase*

**HOW I HAVE APPROCHED THE PROJECT:**

As per the instructions given in the LMS that there were 3 separate data files from which I want to load the dataset and those data files were in the format of **.dat** extension. So, while loading dataset from those files I had to put some extra parameters with function **pd.read\_csv** and those specialparameters are **“separator=::”, “engine= python” and “names=[column\_names]”.** I did the same step for all the 3 datasets.

Then after loading the dataset next task is to merge those 3 datasets based on a particular key and as per the instructions I merged all 3 datasets based on first **key= movieID** and second **key=userID**. And thus my final dataframe was ready for further analysis.

After that I checked the shape of the dataframe and viewed the info for the same. From that noticed that there were around **Ten Lacks data records** present in the dataframe and it was tedious as well as time-consuming to process such kind of huge dataset. So I decided to slice it into **One Lacks record.**

After that I did EDA process in which first I checked for missing values then the main task was to convert the Categorical column “Genres” into dummy form using **One-Hot-Encoding.** For this, first I converted this column into list and then using **split and Set** function separated all the unique Categories into list. Then write a loop for dummy conversion. And then concat this new Genres dataframe and Original dataframe. Then I checked the datatype of all the columns and found “Gender” column was still in Categorical format**.** So, I did **“encoding’** for this column but without creating new columns for each unique Category.

After that I noticed that the columns “**Age” and “Occupation**” was not on the **Standard Scale** compared with other columns. So, I did “**Scaling**” using **Min-Max-Scalar** for **Normalize** those columns and thus made whole dataframe in Standard format. And thus finalized the dataframe into further analysis.

After that I did some **Data Exploration** to gain some insights from the data.