# AIML Feature ENGINEERING Assignment 1

Submission Date: 22nd Dec 2021 - 11:59 PM

Weightage: 15%

The data files Bollywood.csv and Bollywood - 2 .csv contains the box office collection and social media promotion information about the movies released in 2013 – 2015 period.

Following is the information about the columns (names can vary):

* Serial Id of movie
* Release date of movie
* Release time of movie like Long Weekend (3), Festive Season (1), Holiday Season (2), and Normal (4)
* Genre of movie like Romance, Thriller, Action, comedy etc.
* Budget of movie in crores
* Box office collection of movie in crore
* YouTube views of movies trailers
* YouTube likes for the movie trailers
* YouTube dislikes for the movie trailers

Q1. Identify and rectify the eight prominent data quality issues present in the given datasets. [8]

Q2. How many records are present in the combined dataset? Print the metadata information of dataset. [1]

Q3. How many movies in each genre got released in different release times? [1]

Q4. Which month of the year, maximum movie releases are seen? [1]

Q5. Which movies have got more dislikes than likes on Youtube? What are the number of views on these movies? [1]

Q6. Which month of the year typically sees most releases of high budgeted movies, that is, movies with budget of 25 crore or more? [1]

Q7. What is the correlation between Box office collection & Budget? How would you interpret it? [1]

Q8. Which are the top 10 movies with maximum return of investment (ROI)? Calculate ROI as (Box office collection – Budget) / Budget. [1]

References:

1. Chapter 2, “Data”, Introduction to Data Mining eBook: Vipin Kumar, Pang-Ning Tan Michael Steinbach
2. **[Pandas documentation](https://pandas.pydata.org/docs/)**

**Notes:**

* This is a take-home assignment to be carried out by each learner individually and independently.
* This is programming exercises - requiring both datasets to be used – on Jupyter notebook environment.
* You may consult / discuss with other learners peripheral aspects such as the environment but not on solving the specific problems in terms of design or implementation.
* You have to write the appropriate Python code in Jupyter notebook to support you answers and submit with following nomenclature

Final document - FE\_Assignment1\_<Student\_ID>.ipynb

* Strictly do not upload any zip/rar files.
* Provide appropriate justification when processing the data or arriving at the conclusions.
* In case of any further queries, if those are generic once, learners are encouraged to use discussion forums, otherwise they can reach out to me at [murtuza.dahodwala@wilp.bits-pilani.ac.in](mailto:ppawar@wilp.bits-pilani.ac.in).
* Manage your efforts properly as there is no scope to shift the deadlines announced above.