**TERM PROJECT – MOVIEW METADATA ANALYSIS**

For Term project, chosen data set that is associated to Movies Metadata and performed Exploratory Data Analysis, Hidden insights based on available data in data sets, performed correlation analysis on movie reviews and visualize the same to display data graphs.

To mee the objective, below process steps are followed:

* Data Collection
* Data Cleaning
* Data Exploratory Analysis
* Data Modeling
* Data Visualization

*Outcome:*

* Displayed Histogram of movie reviews:

Observations:

* + The dataset seems to be fairly distributed although it does not contain equal information on all score ranges.
  + As can be seen most of the scores are ranged in the [5. 8] range.
  + This might bias the results of some low-end algorithms.
* Displaying Mean IMDB Score Vs Country of movie production:

Observation:

* + At first sight, it seems that the main target variable “IMDB\_Score” is fairly distributed among countries but that is not at all true and can be explained by the next graph.
* Displaying Number of Movies Vs Country of movie production:

Observation:

* + Clearly, it can be seen that USA and UK are the most contributors of movies and hence the scores within this dataset. This makes the dataset heavily biased on country of movie origin.
* Displaying Mean\_IMDB\_Score Vs Rating by Motion Pictures Association

Observation:

* + It is interesting to see how the rating affects the score but just as in the previous case the country or origin was biased, this aspect is also fairly biased. Also, it has too many cases to be incorporated in our analysis even after applying encoding
* Displaying Number of Movies Vs Rating by Motion Pictures Association:

Observation:

* + We can see the actual ratings which actively contribute to the dataset.
* Displaying Scatter plot Imdb\_Score vs Duration of a Movie

Observation:

* + Overall time of a movie seems to be playing a nominal role in deciding the score of a movie.
  + A small trend indicates long movies having better review scores.
* Displaying Scatter plot Imdb\_Score vs Budget of a Movie

Observation:

* + Few data points are acting as outliers and they are in the range of gross earnings of a movie where as this graph is about the budget.
* Displaying Scatter plot Imdb\_Score vs Budget of a Movie After Filtering

Observation:

* + There exists a fair trend and in a way movie scores are being impacted by the budget invested in them.
  + Also, since it is a continuous feature, it hold great potential in acting as a predictor in our models.
* Displaying Scatter plot Imdb\_Score vs Gross collection of a Movie:

Observation:

* + There exists a fair trend and in a way movie scores are being impacted by the gross earning of the movie.
  + There will be caveats for incorporating this in our predictor variables which will be explained during analysis.
* Displaying Number of Movies Vs Year of movie production:

Observation:

* + Clearly, we can see that as the Year progresses, more number of movies are generated. It is quite understandable as the technological advancements would have made it easier to produce more movies.
* Displaying Mean IMDB Score Vs Year of movie production

Observation:

* + The general trend of imdb scores is showing a slowly reducing one as the years progress.
* Displaying Scatter plot Imdb\_Score vs Director Facebook Likes

Observation:

* + We can see that the movies with higher scores are having a good proportion of directors who are popular as judged by their facebook likes.
* Displaying Scatter plot Imdb\_Score vs Actor 1 Facebook Likes

Observation:

* + This field seems to be fairly even along with few outliers. It can act as a good candidate for our predictions.
* Displaying Scatter plot Imdb\_Score vs Actor 2 Facebook Likes

Observation:

* + We can see that actor popularity is indeed affecting movie ratings. We will be considering all actor rating in the analysis.
* Displaying Scatter plot Imdb\_Score vs Actor 3 Facebook Likes

Observation:

* + Another one to bolster our decision. Hope fully they are not correlated and there is no multicollinearity.
* Displaying Count of Movies Vs Number of Faces in Poster

Observation:

* + Most of the movies seem to have fewer faces in the movie poster and we saw from the first plot that most of the scores are in the [5,8] range. So most high rated movies would be having a relatively lower number of faces in their posters. This indicates some relationship and could be used as a predictor variable.
* Displaying Correlation Analysis Plot

Observation:

* + We can easily short list the candidate predictor variables.
  + We are choosing the following after careful analysis:
    - "movie\_fb\_likes"
    - "actor\_1\_fb\_likes"
    - "actor\_2\_fb\_likes"
    - "actor\_3\_fb\_likes"
    - "director\_fb\_likes"
    - "duration"
    - "gross"
    - "budget"
    - "title\_year"
    - "faces\_in\_poster"