**Applied Data Science Assignment 1**

Student ID Number : 22033166

Name: Bhavana Maddukuri

Git hub repository : <https://github.com/bhavanamaddukuri/ads-asssignment-1>

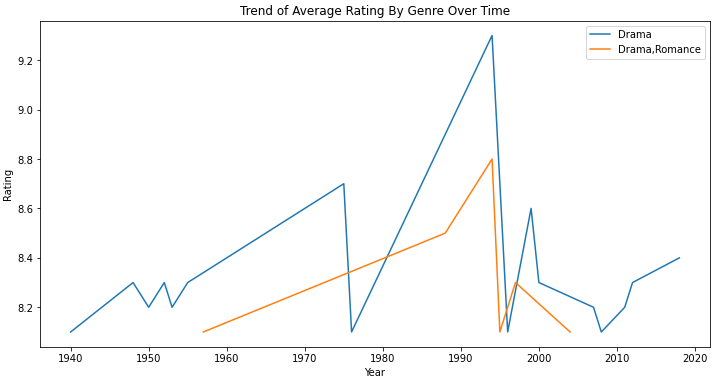
**DATA VISUALIZATIONS**

# Introduction:

In this project we are working with IMDB top 250 movies dataset that is being collected from Kaggle and could be found at <https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset>. Our main goal in this project is to create some meaningful visualizations to identify hidden information from the dataset in this case. In this project, we will create one line plot, one bar chart and one regression plot using pyplot library associated with matplotlib.

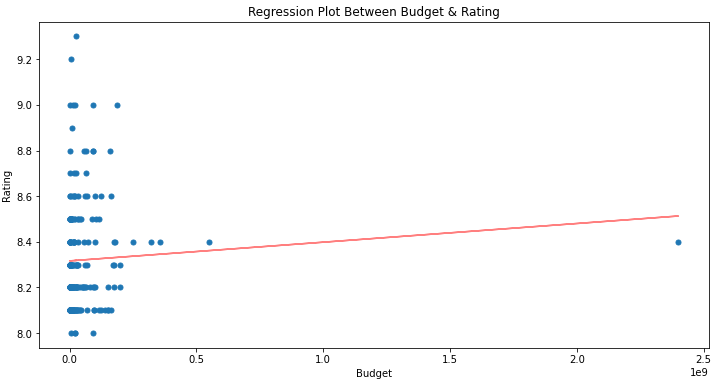
# Data Visualizations:

We have generated a line chart as our initial visualization to display the movie ratings across various genres. After filtering the dataset, we focused solely on the drama and drama & romance genres for this plot. The resulting plot is provided below. The plot highlights that during the period between 1990 and 2000, there were certain years in which the movies had received their highest average rating. Similarly, there were some years during which the movies had received their lowest average rating.



We picked a regression plot here because it is useful for exploring the relationship between two variables. In this case, we have used a regression plot to examine the relationship between two quantitative variables, which are the budget of a movie and the rating. The regression plot can help us to identify any patterns or trends in the data. For example, if there is a positive relationship between the two variables, we would expect to see the points on the plot to generally form a line that slopes upwards from left to right. Alternatively, if there is a negative relationship, the points would generally form a line that slopes downwards from left to right. The regression line on the plot provides a summary of the relationship between the two variables, and can help us to make predictions about future observations. The slope of the line can be interpreted as the average change in the dependent variable (in this case, movie ratings) for a one-unit increase in the independent variable (the budget spent on the movie).

The graph thus created is attached below and we can see a positive relationship between both variables. This means that higher budget of a movie leads to better rating whereas lower budget leads to lower rating of a movie.



Finally, the top 5 writers based on box office value generated of all movies of those writers and the result thus obtained is attached below. A bar plot is an effective way to compare values across different categories. In this case, we used a bar plot to compare the box office values of the top 5 writers in our dataset. The bar plot allows us to quickly identify which writers had the highest box office values, and by how much they outperformed the other writers. Additionally, the plot makes it easy to compare the box office values of the top 5 writers to one another, and to see any patterns or trends in the data.

Now from the graph attached below, we can see that the writer at the top based on box office value is Cristopher Marcus, the second highest value is of JRR in this case. Thus this shows us that the writers who made the highest box office values are Cristopher and JRR.

