# **Data Science Case Study – Version 1**

Analyze the attached dataset ‘movie\_metadata’ and identify the insights which can be generated from this data. The dataset has 28 variables which are explained in detail in the ‘variable description’ tab.

**Key Questions to be answered:**

1. What are your observations based on exploration of this data?

In data exploration phase I created various plots to visualize the correlation and general trends in data. The data exploration phase revolves around some basic points namely:

a) Non-linearity

b) Correlation of error terms

c) Non Constant Variance

d) Outliers

e) Collinearity.

I observed the following things:

1. The data has some bias in terms of the countries that the movies belongs too, Color of the film, language, content rating and the release year
2. The movie score follows a decreasing pattern with newly released movies rated less compared to the older movies.
3. As the data was collected from general population the movie rating is supposed to follow a normal distribution but the ratings are on the higher side with mean roughly around 6.5
4. The data has lot of missing values in terms of gross income which cannot be imputed thus significantly reducing the size of available training data
5. I also created word cloud for the movie genres and found that drama is the most frequent one with the top five being Drama, Comedy, Thriller, Action, Romance
6. Regarding the important plot keywords, the five most important are Love, Friend, Murder, Death and police.
7. The data also has some useless categorical variable and also some leaky features like gross of a movie, Facebook likes for the movie as most of the data is not available for new movie that is about to be released or just released before few days so we should be cautious before using such information
8. I also took help of QQ plot, Heatmap and Pearson’s correlation for appropriate use.
9. What is the recipe to make a blockbuster, profitable movie? Share your hypothesis and insights based on the data here.

For making a blockbuster and profitable movie the following are the key steps I followed for coming to the conclusion:

1. I assumed that the problem can be considered as a regression problem as the output variable is continuous
2. For selecting the data points, I assumed that a movie may be considered block buster if score is above 7.5 and the average profit is 2 times the mean profit value
3. The refined data set was split into test train data using sklearn
4. The numeric features were considered and the profit was considered as desired output variable.
5. 2 machine learning algorithms were implemented namely Multi linear regression and Random Forest Regression.
6. The features were sorted by the importance and are represented in the figure below.
7. Thus, it was observed that top five feature for having a profitable blockbuster movie are num\_voted\_users, num\_user\_for\_reviews, title year, budget and num\_critics\_for\_review

