**Introduction**:

Film industry is one of the top grossing industries in the world. Millions of dollars are invested in the making of each movie expecting high margin of profit. The primary goal by undertaking this project is to investigate the influential factors affecting the success or failure of a movie which could be game changing not only for the producers but also the audience. With all this in our mind we set on a quest to answer this question: Whether is it possible to use machine learning algorithm to predict if movie will be a success (generate at least as much revenue as its budget) or a flop?

The primary source of our data set is [Kaggle](https://www.kaggle.com/rounakbanik/the-movies-dataset) but some columns were added by parsing data made publicly available on [IMDb’s website](https://datasets.imdbws.com/). Following are the modifications that we applied to our raw data set.

**Data Dictionary:**

*Description of each column goes here.*

**Data Cleaning:**

The first operation performed was to remove all those columns from our data set which were irrelevant e.g. “homepage” and “poster\_path”. Then we searched for corrupt values in the columns e.g. “budget” column contained alpha-numerical values so we removed all such rows.

Two variables “Genre” and “Production\_Companies” were in JSON format in raw data set so one of the challenges we faced during cleansing data was extracting and converting them to string data type. “release\_date” column was also converted to its proper datetime format using pandas. Based on the months in this column, we created a new column because we believed that movies released in vacation months e.g. summers or at Christmas/New Year time have a higher chance of getting good profits.

Since the dataset from Kaggle was two years old and “average\_rating” and “vote\_count” are values which keep updating on very regular basis as more people watch and rate the same movies so we got those as well as the “Director” names from the largest and most authentic movie source i.e. IMDb.

To keep the outliers in our data set to a minimum we excluded all those rows for which budget value was less than $100,000 and revenue value less than $1000. Afterwards we created our target column by dividing revenue by budget. If the value obtained was greater than 1, we categorized that movie as “success” else “flop”.

Last step performed was to remove all duplicate values. We started with a dataset containing 45,000+ rows but after cleaning ended up with having only 2,222.

**Exploratory Data Analysis:**

*Graphs to be added later*

Dependent Variable:

First lets take a look at our target variable. From the histogram below we see that there are approximately \_\_\_ successful & \_\_\_\_ flop movies in our cleaned data frame.

Independent Variable:

Out of 2,222 movies the ratio of Revenue/Budget (“status” column) that they generate lies between \_\_\_\_ & \_\_\_\_. Most of the data is concentrated around \_\_\_\_\_.

Coming towards the release year data, we found out that the oldest movie in our dataframe was released in \_\_\_\_ whereas the most recent one was released in \_\_\_\_\_\_ (*Kaggle dataset we are using is 2 years old).*

The distribution of “runtime” depicts that most movies are concentrated around \_\_\_\_\_\_ minutes.

Then we plotted histograms for “VoteCount” & “VoteAverage”. Movie that got the maximum number of votes and rating was “The Shawshank Redemption”.

Although there were 18 genres in our data set but most of the movies fell under the category of Action, Drama and Comedy.

Finally we plotted the correlation heatmap between numerical variables. The only figure that is worth mentioning is 65% of positive correlation between budget & revenue. But overall we see no significant relationship between any of our numerical features.

**Modeling:**

**Conclusion:**

**References:**