# TMDB Dataset Analysis Report

**Introduction**: This project investigates the movies dataset containing 10,000 rows and its related parameters with the use of the following libraries:

- NumPy
- Pandas
- MatplotLib

**Questions** that needs investigations are:

- Which genre has maximum releases, and which has minimum releases?
- Which 5 movies have grossed the highest profit in the period 2010-2019?
- How is the revenue associated with popularity?
- Maximum number of movies released in which year?

## Procedure:

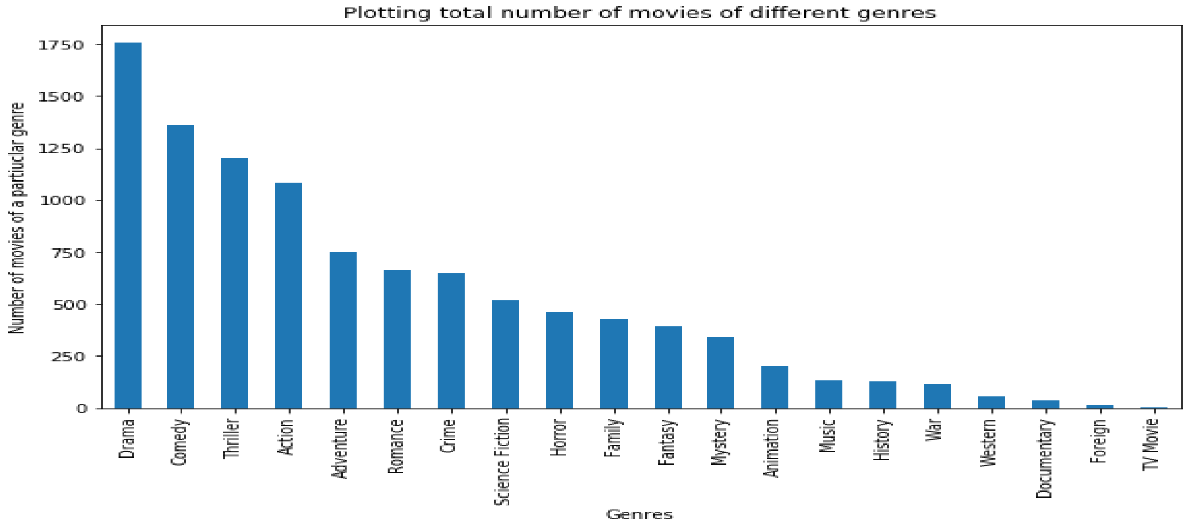
Steps taken in the analysis of this dataset are as follows:

1. Read the csv related to this dataset using pandas and found that there are unnecessary columns present. These columns are unnecessary as they don't much contribute to the questions that I want to explore.
2. Dropped columns such as 'overview', 'homepage', 'tagline', 'keywords', 'imdb_id', 'budget_adj', 'revenue_adj'
3. Also, observed using dtypes( ) method that release_date is in String format, so converted it using to_datetime( ) method of pandas.
4. In next step, I started checking for any null values being present in my data and found out that there were a huge number of such values in various cells. Took basically two steps for handling null values which are as follows:
    - Any cells that had null values present, filled it out with 0 using fillna( ) method.
    - Removed any rows containing either containing revenue or budget as 0 as it would impact my analysis for the second question.
5. Next, I checked for any duplicate rows present in my dataset, where I found one row being duplicated so removed it using drop_duplicates( ) method.
6. After all the cleaning till this point, the number of rows and columns in my dataset were 3854 and 14 respectively.
7. Next, I saved the cleaned datasheet to a new file, which I used for reference while making any exploration beyond this point.
8. For the question No. 1, I took help from udacity forum where it was advised that I make a function to separate all the values in the cells containing separator '| '. This function takes an argument with column name and checks for values in the rows and split the strings and return the count of the values.
9. For question No. 2, created a new column named profit and assigned it a value after deleting each row's revenue by budget. After the new returned dataset, I performed a query to check the release_year between 2010 and 2019(both including), then used nlargest( ) to find the top 5 rows from that dataset with maximum profit and plotted the graph.
10. For question No. 3, plotted a graph between revenue and popularity just to see how it varies and found out that it is proportional.
11. For question No. 4, first found out the number of times a year has been repeated in the column 'release_year' by using values_counts() method on 'release_year' column, and used that count to plot the graph between number of movies released in a particular year.

## Conclusions:

Following conclusions can be made based on the data given and the plots drawn:

1.

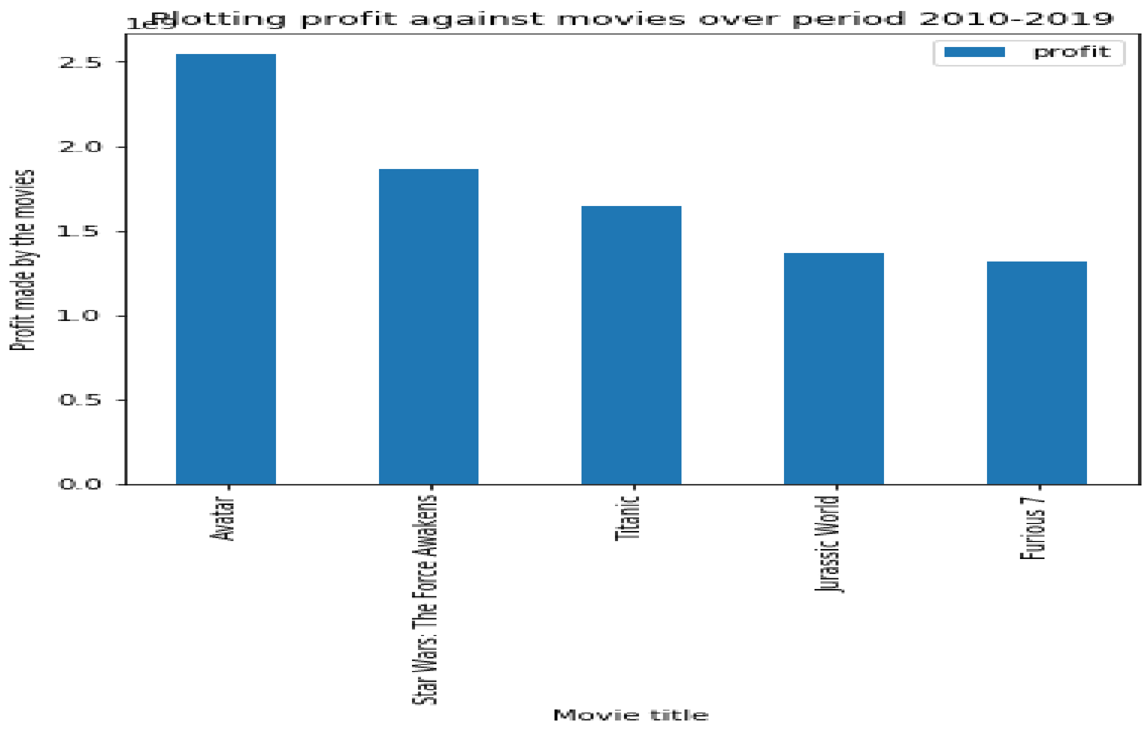


Based on the above bar graph, we could see that:

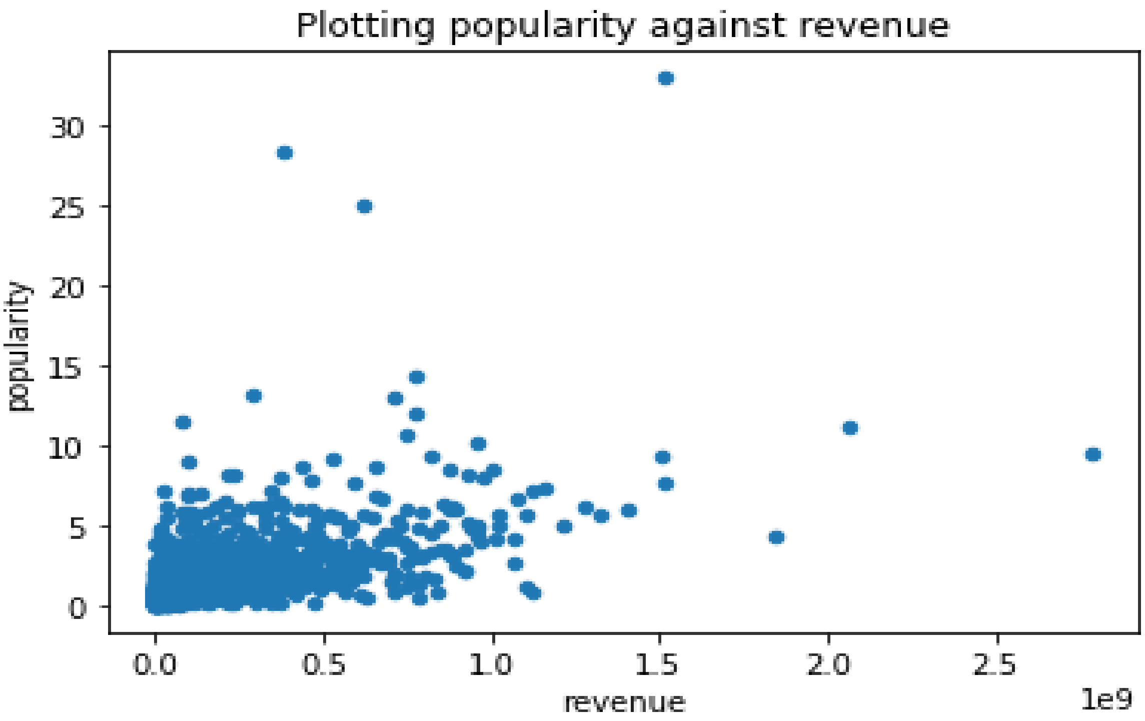
- Top genres of which movies are released are drama, comedy, thriller and action.
- Number of movies released of genre romance and crime are approximately the same.
- Number of movies released of genre history and war are approximately the same.
- Movies with 'drama' as genre has been released in maximum number and movies with 'TV Movie' as genre has been released in minimum number.

2. Based on the below bar graph below, we could see that:

- Profit made by Jurassic Park and Furious 7 are nearly equal.
- Clearly, the top 5 movies with highest profits are in the order:
    - Avatar
    - Star Wars: The Force Awakens
    - Titanic
    - Jurassic World
    - Furious 7



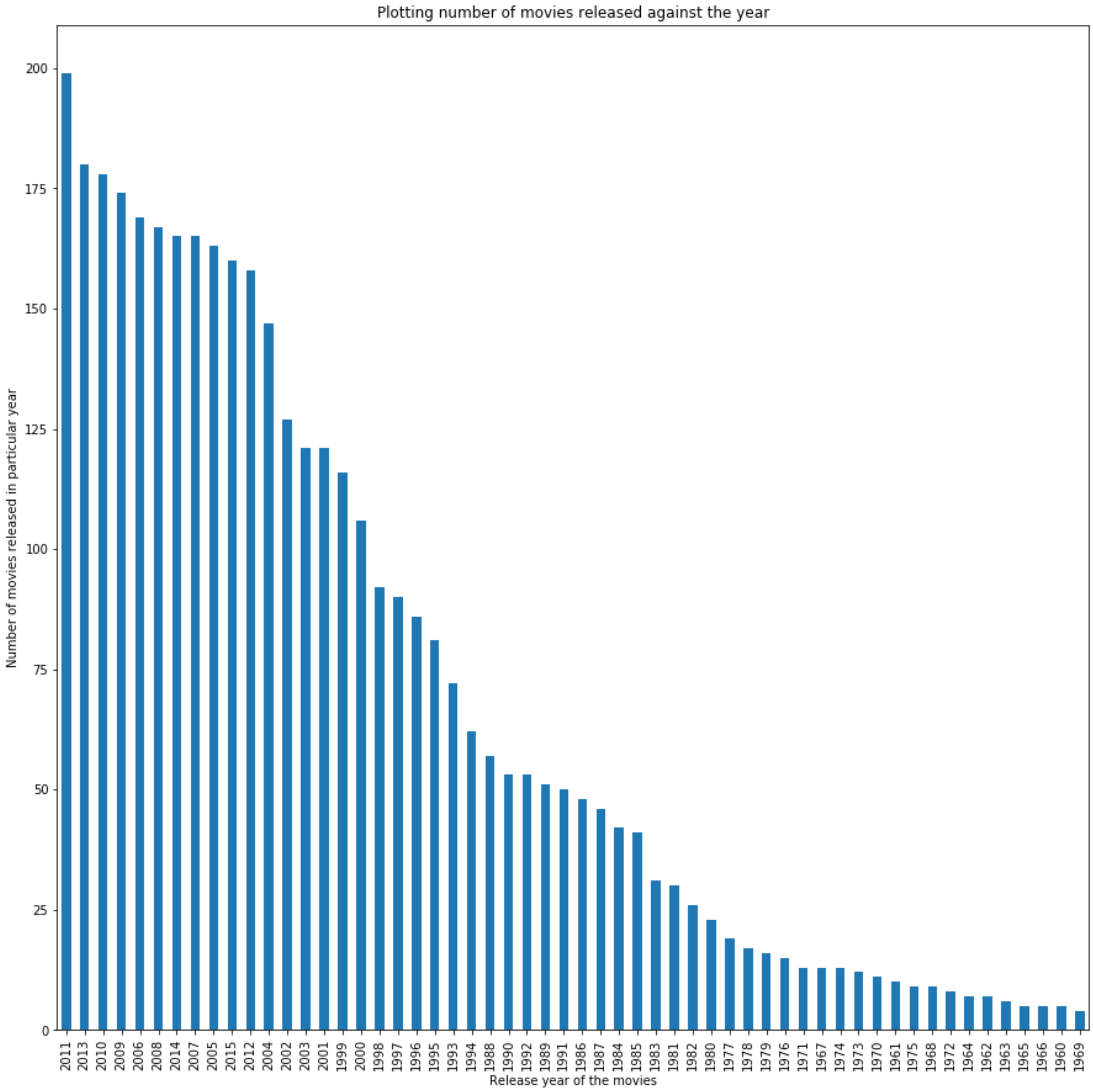
3.



Based on the above scatter graph, we could see that:

- Most of the movies with lesser popularity have generated lesser revenues.
- Most of the movies have around zero popularity.
- With the increase in the popularity, the revenue seems to be increasing.

4.



Based on the above bar graph, we could see that:

- Number of movies released in the period 60s is way less that the movies released in 2010s.
- As per the observation, maximum number of movies were released in the year 2011 with the count of movies almost near to 200.

## Limitations:

The whole analysis was done considering that the revenue and budget are in same currency format, in any other case, the results might vary.