PROJECT ANSWERS:

INVESTIGATE A DATASET(TMDb_Movies Dataset)

Overview:

To complete my project I am using TMDb movies dataset.

This data set contains information about 10 thousand movies collected from The Movie Database (TMDb), including user ratings and revenue. It consist of 21 columns such as imdb_id, revenue, budget, vote_count etc.

QUESTIONS that can analyised from this data set:

- *Movies which had most and least profit*

ANS: Maximum profit earned: 2544505847 (Avatar)

Minimum Profit made:-413912431 (The Warrior's Way)

- *Movies with largest and lowest budgets*

ANS: Maximum Budget: 425000000 (The Warrior's Way)

Minimum Budget: 1 (Lost and Found)

- *Movies with most and least earned revenue revenue:*

ANS: Maximum revenue: 2781505847 (Revenue)

Minimum revenue: 2 (Shattered Glass)

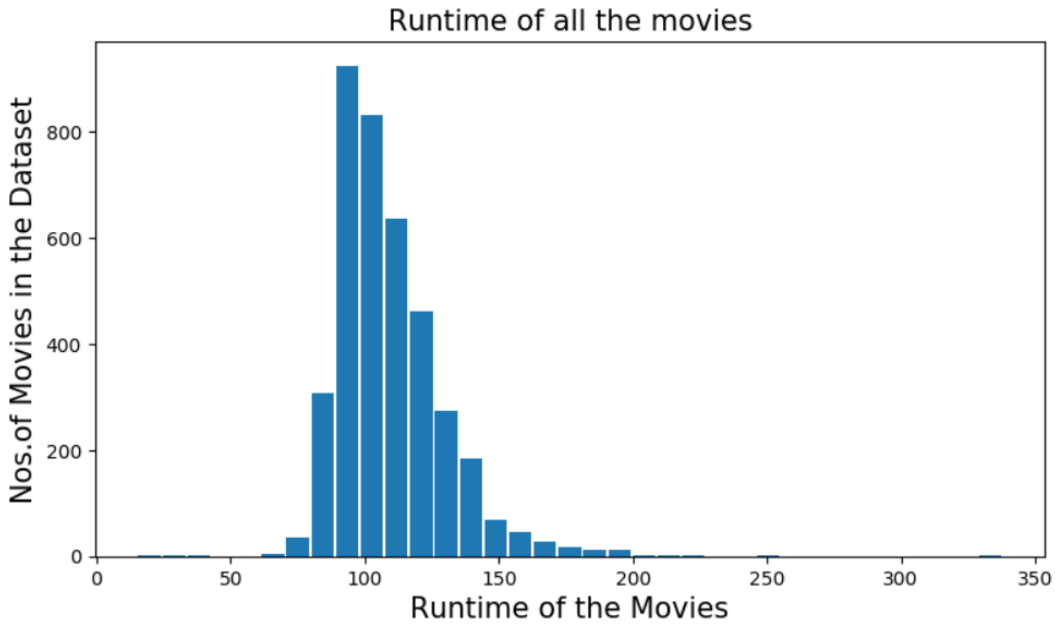
- *Movies with longest and shortest runtime values*

ANS: Longest Runtime: 338 minutes (Carlos)

Shortest runtime: 15 minutes (Kid's story)

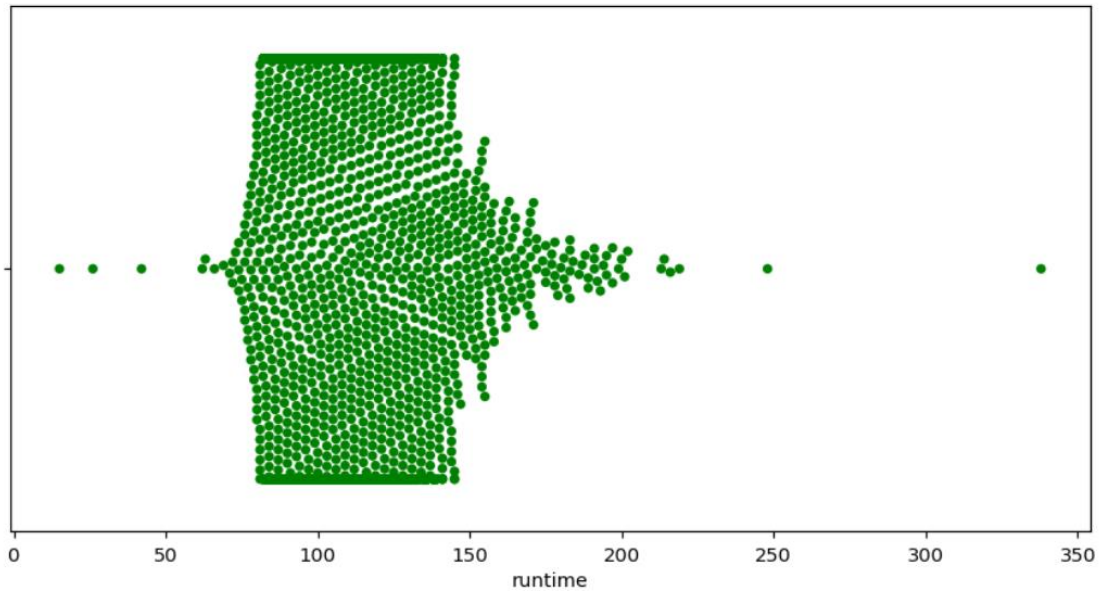
- *Average runtime of all the movies.*

ANS: The average runtime is 109.22 minutes, for a "graphical representation";



The distribution of the above formed graph is "positively skewed or right skewed". Most of the movies are timed between 80 to 115 minutes. Almost 1000 and more no. of movies fall in this criterion.

Another plot indicating exact runtimes:



By looking at both the plot and calculations, we can conclude that:

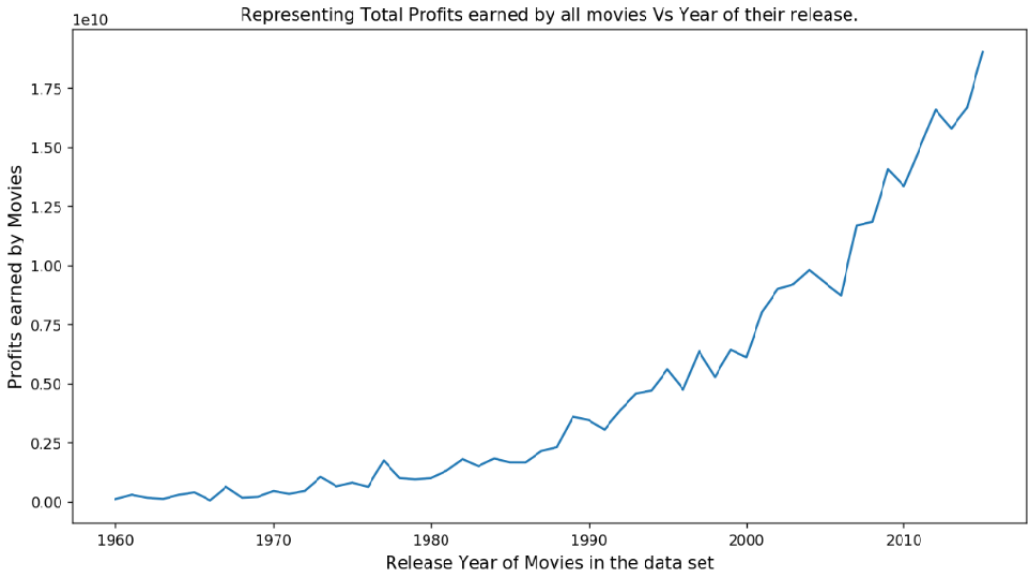
i. 25% of movies have a runtime of less than 95 minutes

ii. 50% of movies have a runtime of less than 109 minutes. (median)

iii. 75% of movies have a runtime of less than 119 minutes

- *In which year we had most no.of profitable movies.*

ANS:



From the plot we can conclude that "2015" had the highest profits.

Before moving further, we need to clean our data again. We will be considering only those movies who have earned a significant amount of profit.

The value will be 75 million!

Therefor our dataset has reduced to 1028 from 3853!

- *Successful genres (with respect to the profitable movies).*

ANS:

i. Comedy 377

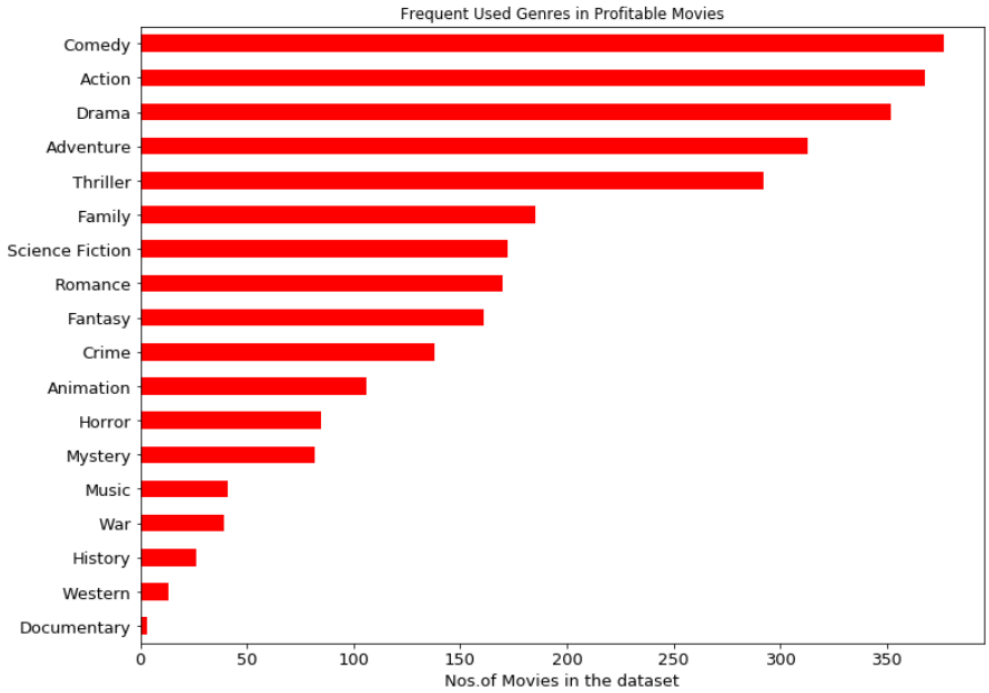
ii. Action 368

iii. Drama 352

iv. Adventure 313

v. Thriller 292

Now for a graphical representation:



- *Most frequent cast (with respect to the profitable movies)*

ANS:

i. Tom Cruise 26

ii. Tom Hanks 20

iii. Sylvester Stallone 19

iv. Brad Pitt 19

v. Cameron Diaz 18

- *Average budget (with respect to the profitable movies)*

ANS: 67.6 million is the average budget when the profit is 75 million, whereas when taken from a different value such as a 50 million profit the budget is 60 million.

- *Average revenue (with respect to the profitable movies)*

ANS: 302.3 million is the average revenue.

- *Average duration of the movie (with respect to the profitable movies)*

ANS: 114.9 million was the average duration of the movie.

Conclusions:

This was a very interesting data analysis. We came out with some very interesting facts about movies. After this analysis we can conclude following:

For a Movie to be in successful or a blockbuster:

i. Average Budget must be around 60 million dollars

ii. Average duration of the movie must be 113 minutes

iii. Any one of these should be in the cast: Tom Cruise, Brad Pitt, Tom Hanks, Sylvester Stallone, Cameron Diaz

iv. Genre must be: Action, Adventure, Thriller, Comedy, Drama.

v. By doing all this the movie might be one of the hits and hence can earn an average revenue of around 255 million dollars.

Limitations: This analysis was done considering the movies which had a significant amount of profit of around 50 million dollars. This might not be completely error free but by following these suggestions one can increase the probability of a movie to become a hit. Moreover, we are not sure if the data provided to us is completly correct and up-to-date. As mentioned before the budget and revenue column do not have currency unit, it might be possible different movies have budget in different currency according to the country they are produce in. So a disparity arises here which can state the complete analysis wrong. Dropping the rows with missing values also affected the overall analysis.

All data values are from the given dataset which have calculated on the jupyter notebook with the help of the various libraries and the code.