Data analysis of IMDB dataset having 50K movie reviews using **Tableau**

(Inhouse Project)

RA1911033010016 - Sanjana N B
RA1911033010021 - Venkata Naga Sai Ram Nomula
RA1911033010034 - Ishwarya M

Abstract

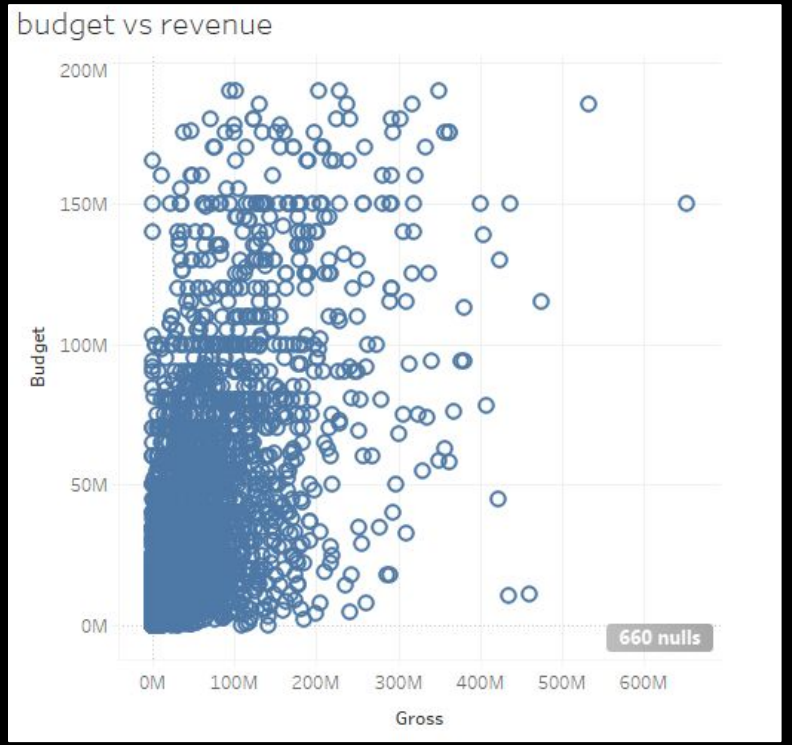
Every year, the global film industry produces thousands of movies with budget of the order of hundreds of millions of dollars. Analysis of movie reviews on social media platforms like Facebook on the basis of genres provides a fair understanding of customer preferences. The prediction of a movie box office and the key factors influencing its success are of great importance to the industry. Machine learning algorithms and Linear Regression models are widely used to make decisions. A detailed study of the outcomes can help us to determine a rough estimate of the reach of the movie among its audience.

Introduction

- In today's world, films with good story content or eye-catching scenes earn over a billion dollars. So the main thing which contributes to the success here is promotions and advertisements.
- Many production industries face losses due to advertisements or because of adjusting the release date of their movies so as to gain maximum profit. They could use the predictions to know when the market is dull and when it is not.
- This scenario shows that we need a software that solves the problem. Techniques such as sentiment analysis have been used in the past. But none of the studies thus far have succeeded in suggesting a model.

- So we are trying to solve the issue using IMDB data to predict the gross and rating of the movie. There is a strong relation between gross and budget .

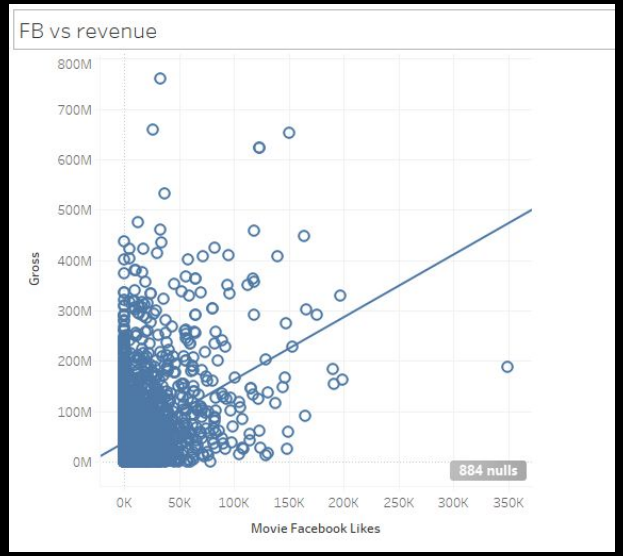
- This shows that the gross revenue is less than $50 million for majority of the movies. Average revenue for the 1000 movies comes close to $83 million.



LINEAR REGRESSION MODEL

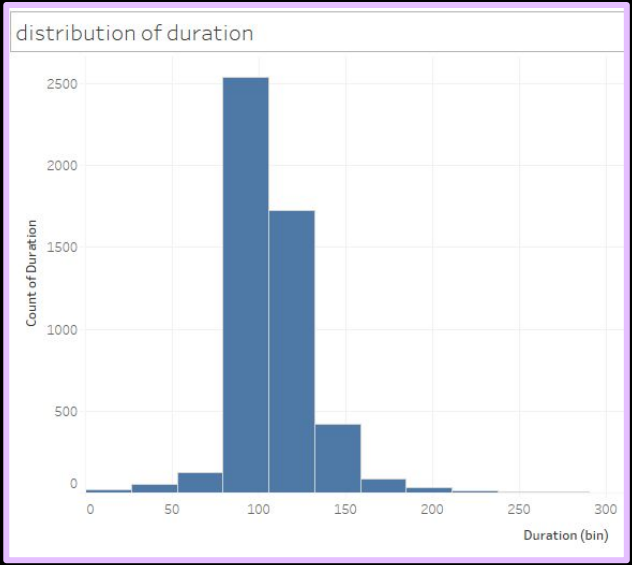
- To analyze each and every factor which can influence the IMDB ratings, so that we can predict better results.
- We can see maximum number of outliers in the drama genre but most significant outliers in crime genre, music genre seems consistent .
- The relationship between imdb revenue and Facebook likes is associated as Linear regression.

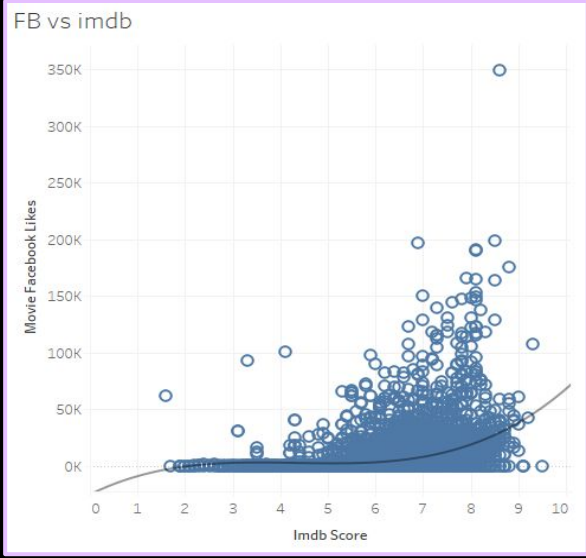
```
LinearRegression()

[ ] model.score(x, y)

0.007359468999322494
```

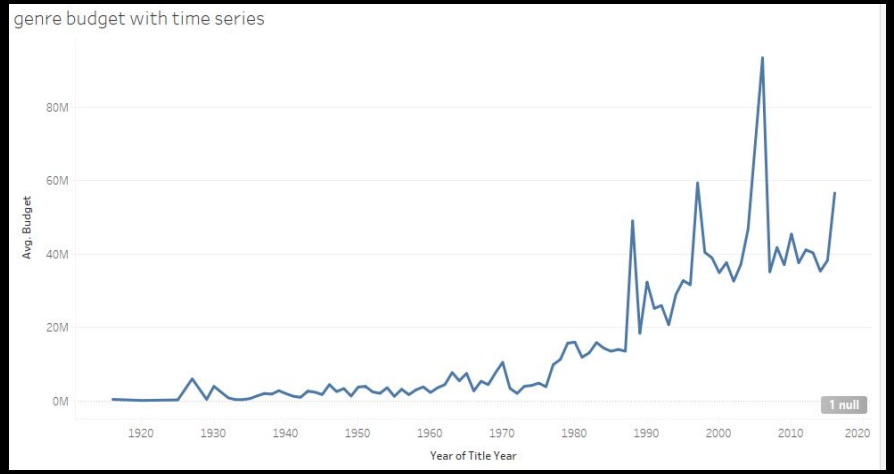


Budget of IMDB movies across the world

- Duration histogram has skewness towards right , we can see that most of the movies appear to be sharply 2 hours long.
- IMDB scores affect the facebook likes as it shows polynomial trend line.







We have 2 columns Gross that indicates how much the movie made in dollars and Budget , the line chat show it spikes from year 1990.

Inferences

- We can see that the top 1000 IMDb voters prefer Sci-Fi over Drama and also the dataset contains more movies from Drama as compared to other genres.
- Relationship between rating and number of votes - we can infer that there is moderately positive correlation, which means movies with higher ratings attract more number of votes.

Conclusion

Analysis of the movie dataset shows that majority of the movies have runtime between 90 and 120 minutes. We also saw that ratings lie between 6 and 7 with mean value of 6.72. While majority of the movies have received less than 100,000 votes, the gross revenue is less than $50 million for close to half of the movies. Since 2007, the average of rating has gone down and it was lowest in 2016. Top genres are drama, action, comedy, adventure and thriller. Although we couldn't establish acceptable correlation between rating and revenue, we were able to establish moderate positive correlation between rating and vote.

THANK YOU!