



IMDb

DATA ANALYSIS OF IMDB DATASET



DONE BY:

RA1911033010016 - SANJANA N B

RA1911033010021 - VENKATA NAGA SAI RAM NOMULA

RA1911033010034 - ISHWARYA M

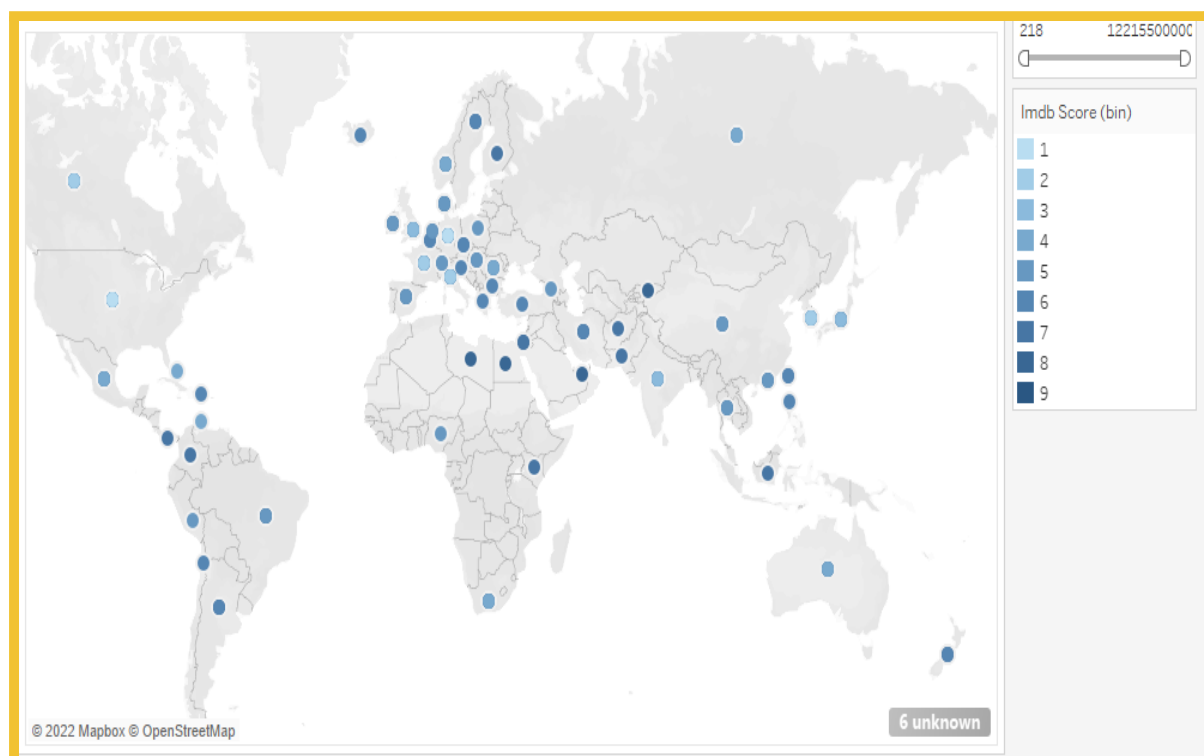


INTRODUCTION

In this modern era, movies with desirable narrative content or fascinating scenes earn over a billion dollars. So the principal aspect which contributes to the fulfillment right here is promotions and classified advertisements.

A number of manufacturing industries face losses because of commercials or due to adjustment in the release date of their films to be able to benefit most profit. They ought to use the predictions to recognize while the marketplace is dull and while it's far not.

This situation indicates that we are in need of softwares that can solve this problem. Techniques consisting of sentiment evaluation had been used withinside the past. But none of the research so far has succeeded in suggesting a model.



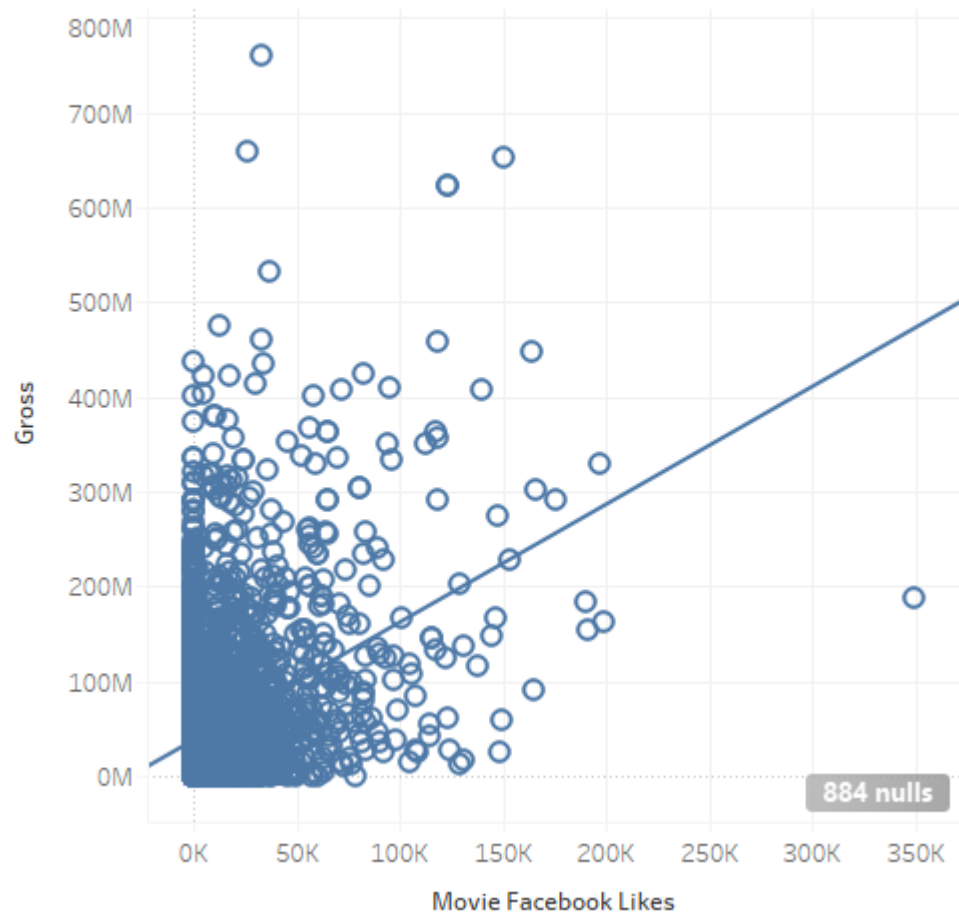
Linear Regression

Finding a line that best fits the data points available on the plot

- 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.



FB vs revenue



```
LinearRegression()
```

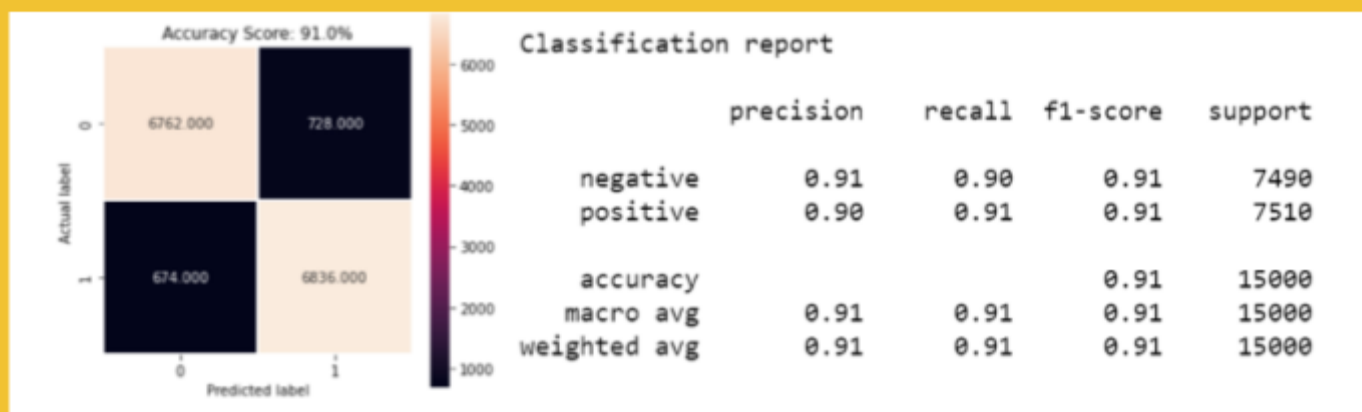
```
[ ] model.score(x, y)
```

```
0.007359468999322494
```

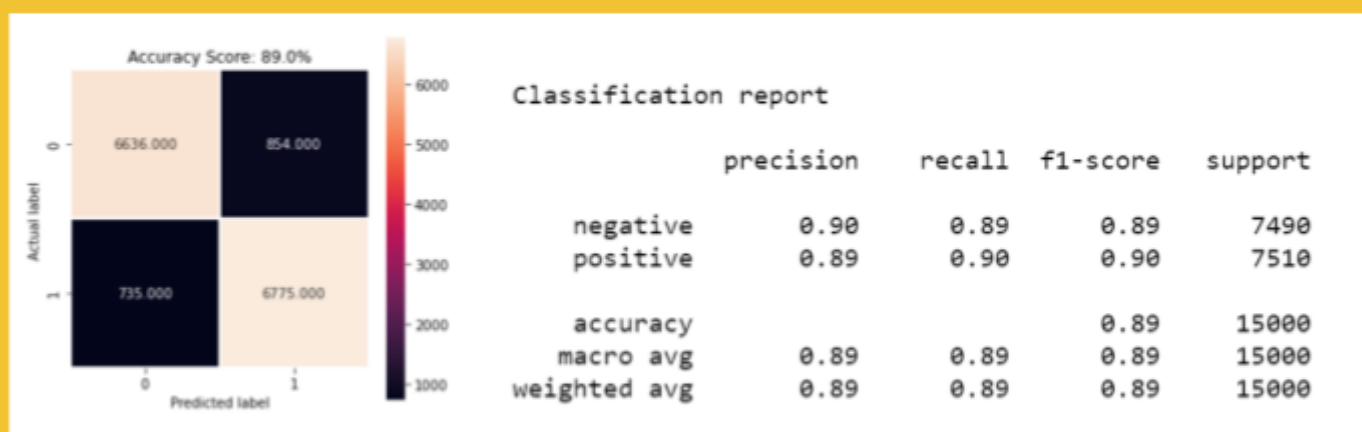
Logistic Regression

Used to calculate the probability of a binary event occurring, and to deal with issues of classification.

Confusion Matrix & Classification Report of Logistic Regression using BoW Features



Confusion Matrix & Classification Report of Logistic Regression using TF-IDF Features



SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is perhaps one of the most popular applications of natural language processing and text analytics. It typically seems to work best on subjective text, where people express opinions, feelings, and their mood. From a real-world industry standpoint, sentiment analysis is widely used to analyze feedback, social media data, and reviews for movies and many more.

The key idea is to use techniques from text analytics, NLP, Machine Learning, and linguistics to extract important information or data points from unstructured text. This in turn can help us derive qualitative outputs like the overall sentiment being on a positive, neutral, or negative scale and quantitative outputs like the sentiment polarity, subjectivity, and objectivity proportions.

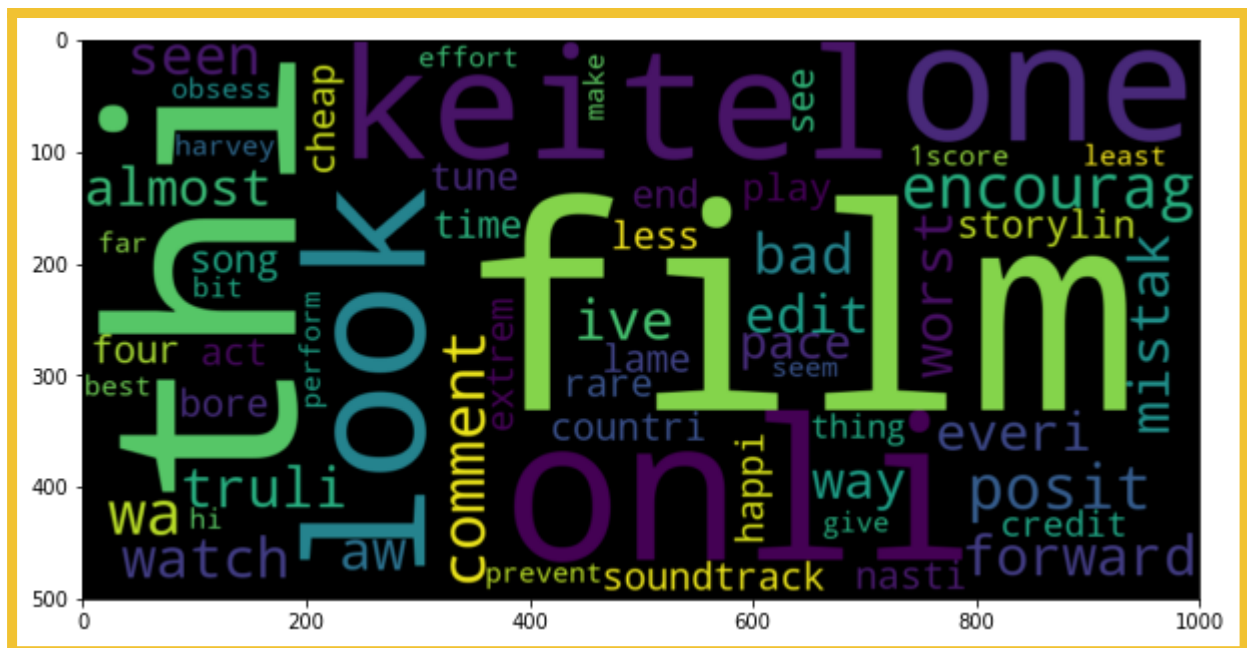
```
#word cloud for positive review words
plt.figure(figsize=(10,10))
positive_text=norm_train_reviews[1]
WC=WordCloud(width=1000,height=500,max_words=500,min_font_size=5)
positive_words=WC.generate(positive_text)
plt.imshow(positive_words,interpolation='bilinear')
plt.show
```

```
<function matplotlib.pyplot.show(*args, **kw)>
```


Word cloud for positive review words

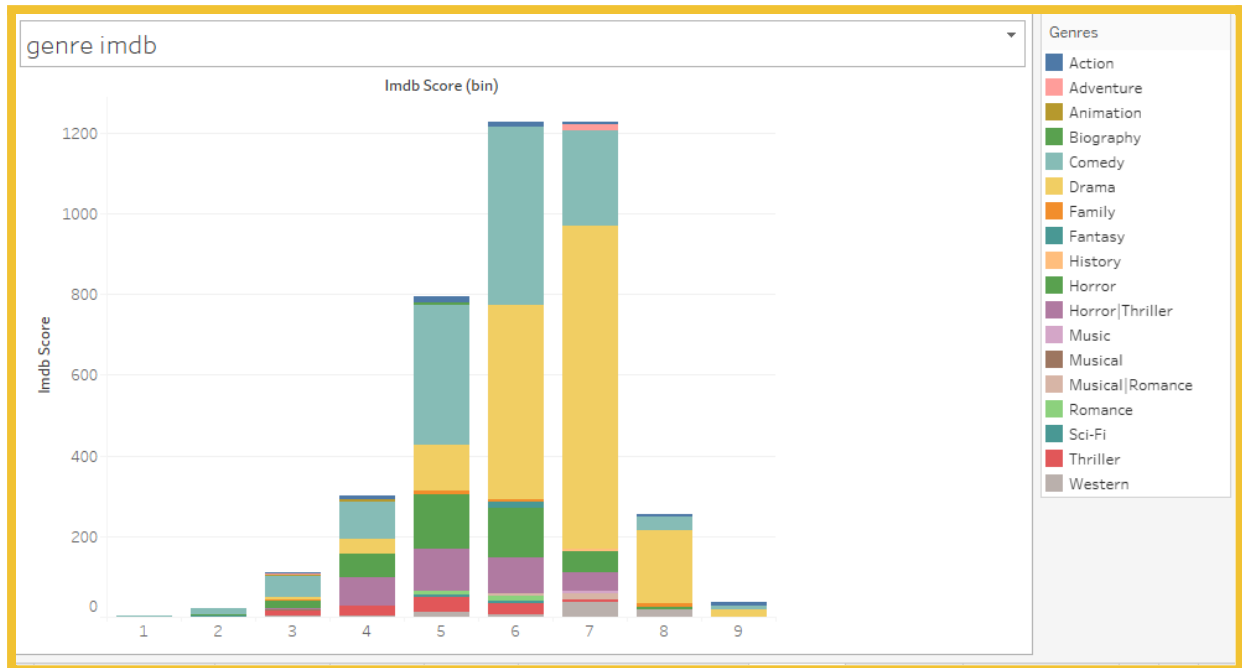


Word cloud for negative review words

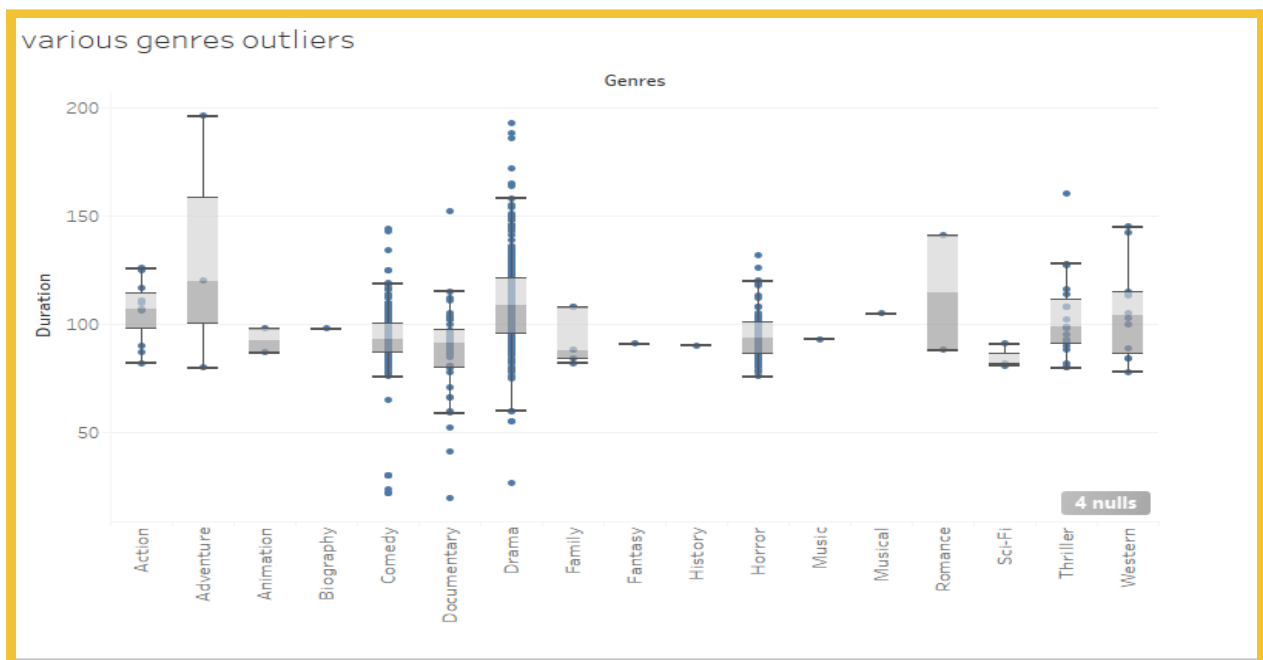


INSIGHTS

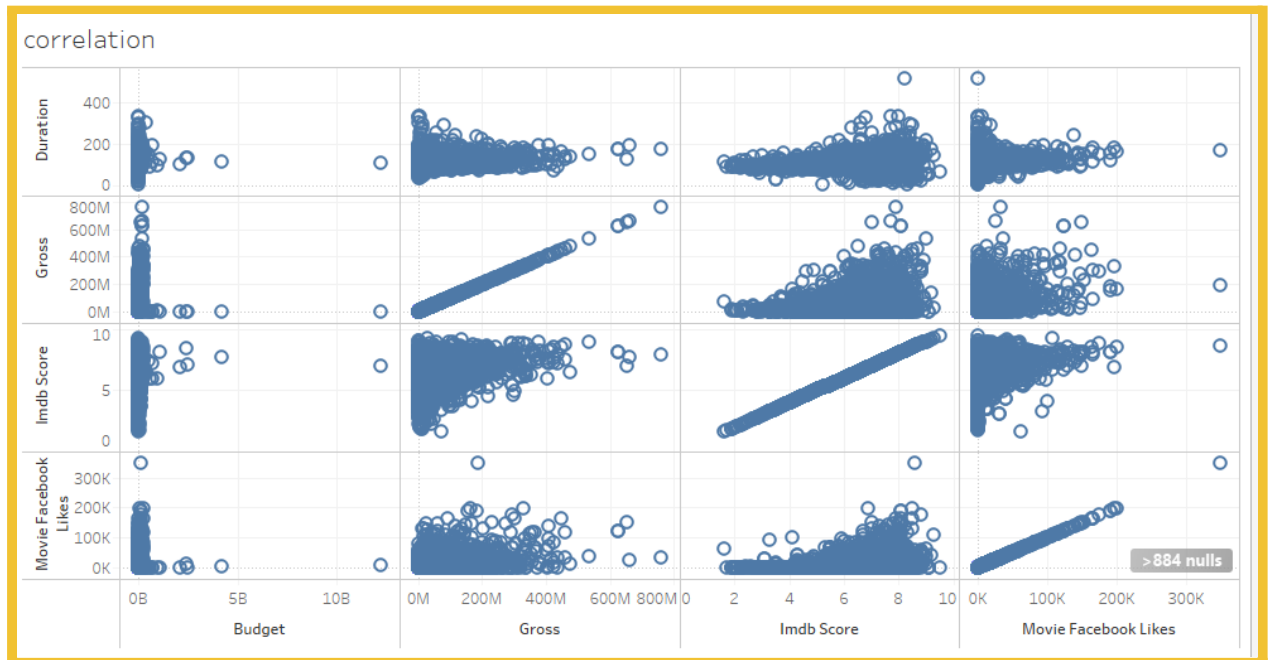
- Drama, action, comedy, adventure and thriller have a fair number of movies. There are the least number of movies in the musical genre.



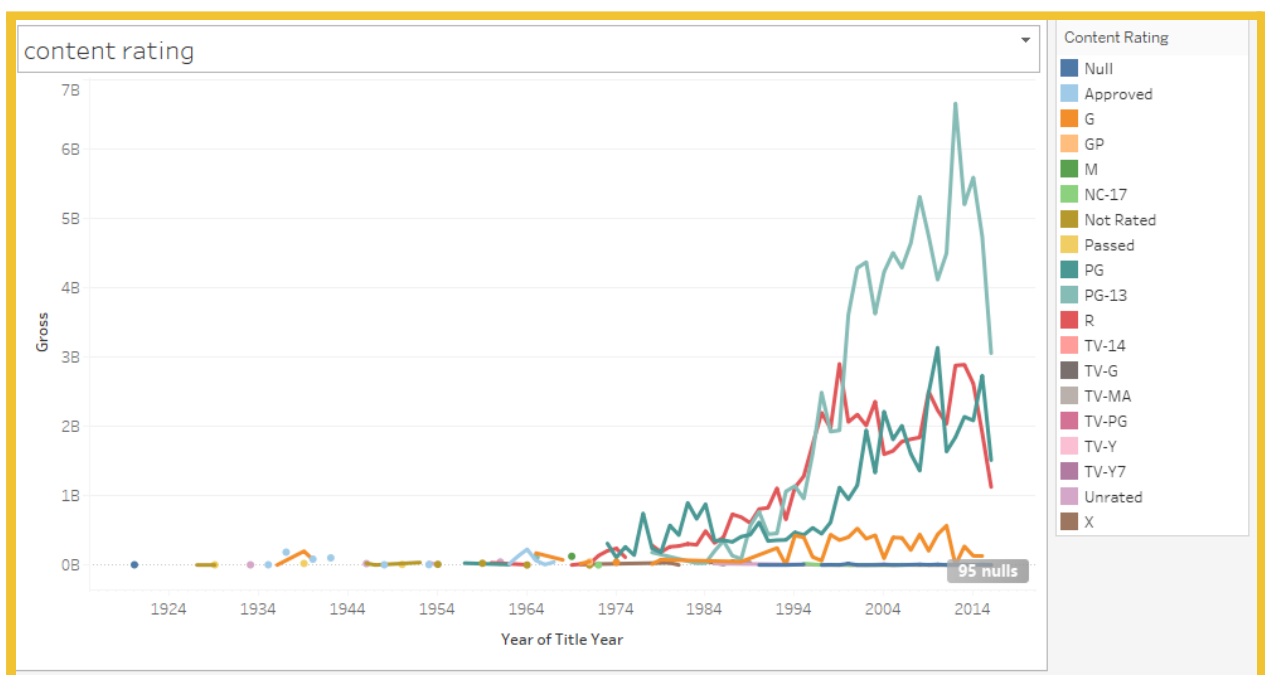
- 'Horror' genre has least ratings and 'drama' has received highest ratings.



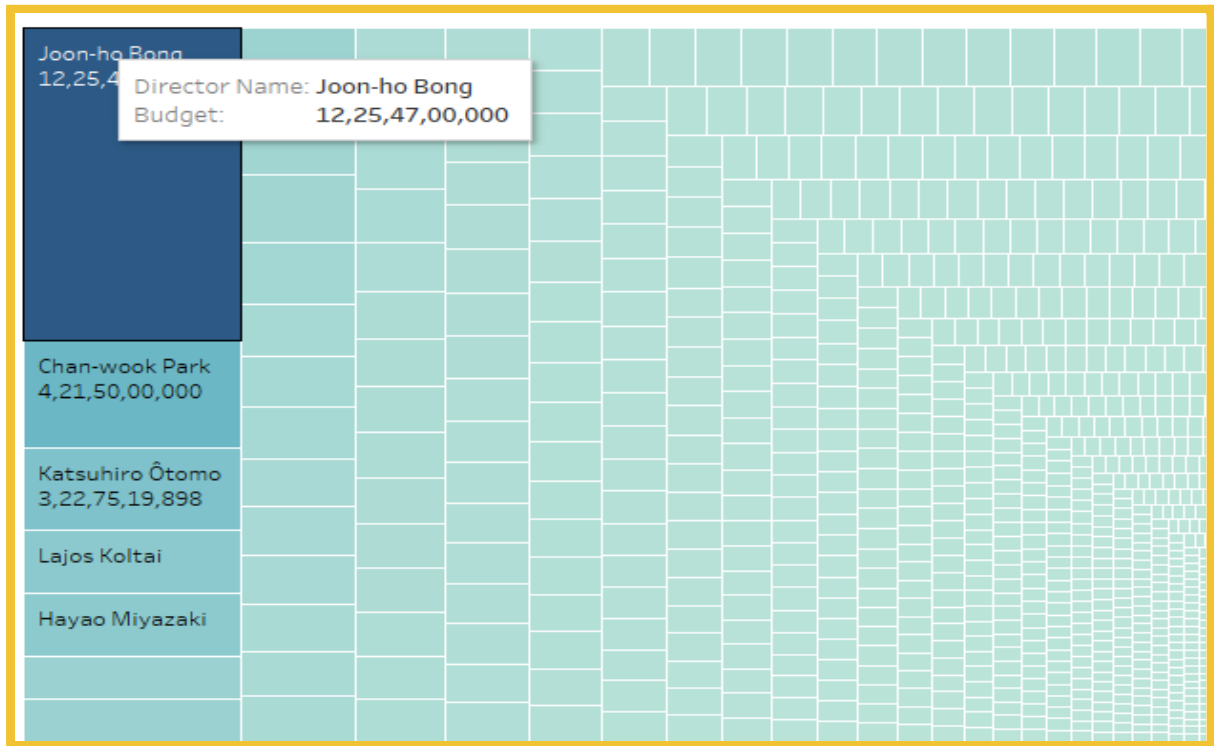
- Plot of rating and revenue will help us understand if movies with better ratings generate higher revenue. Although the correlation coefficient is positive, the relationship is very weak.



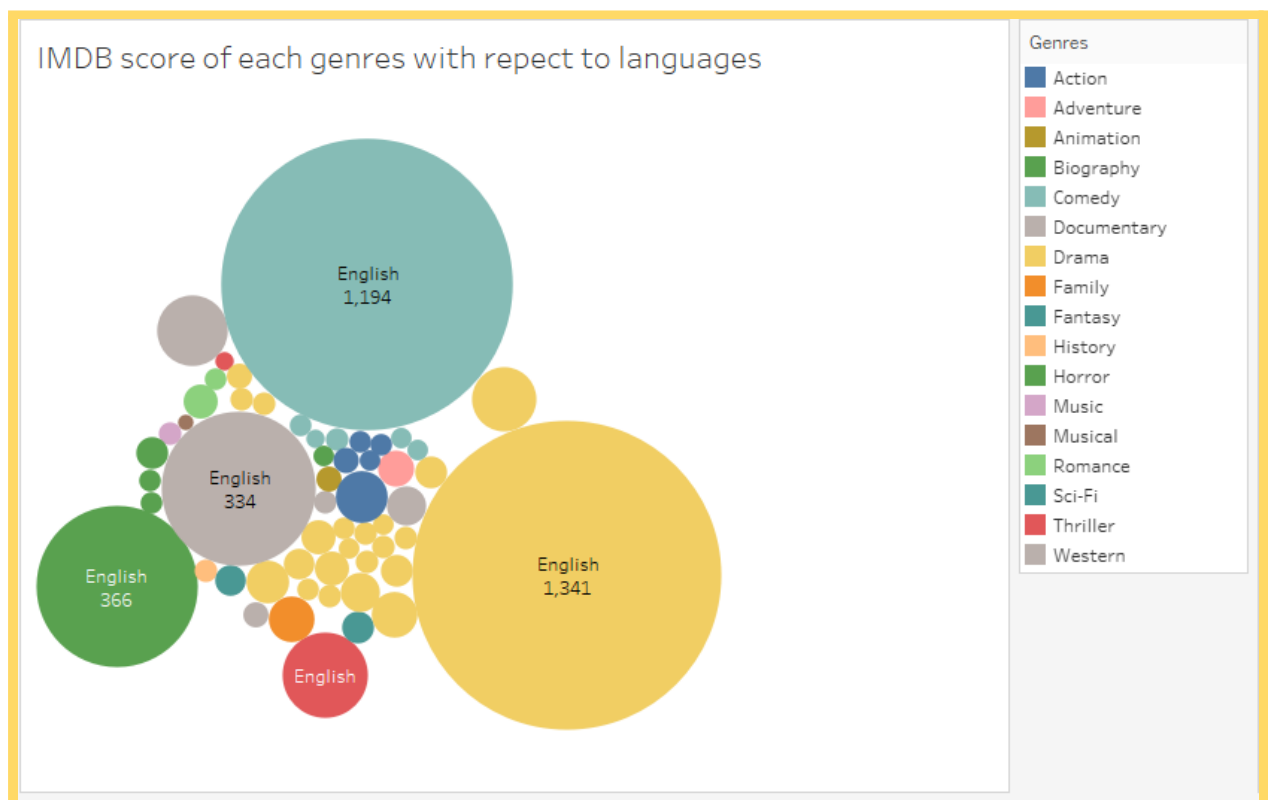
- The attributes field is used to provide additional information on the certificate.



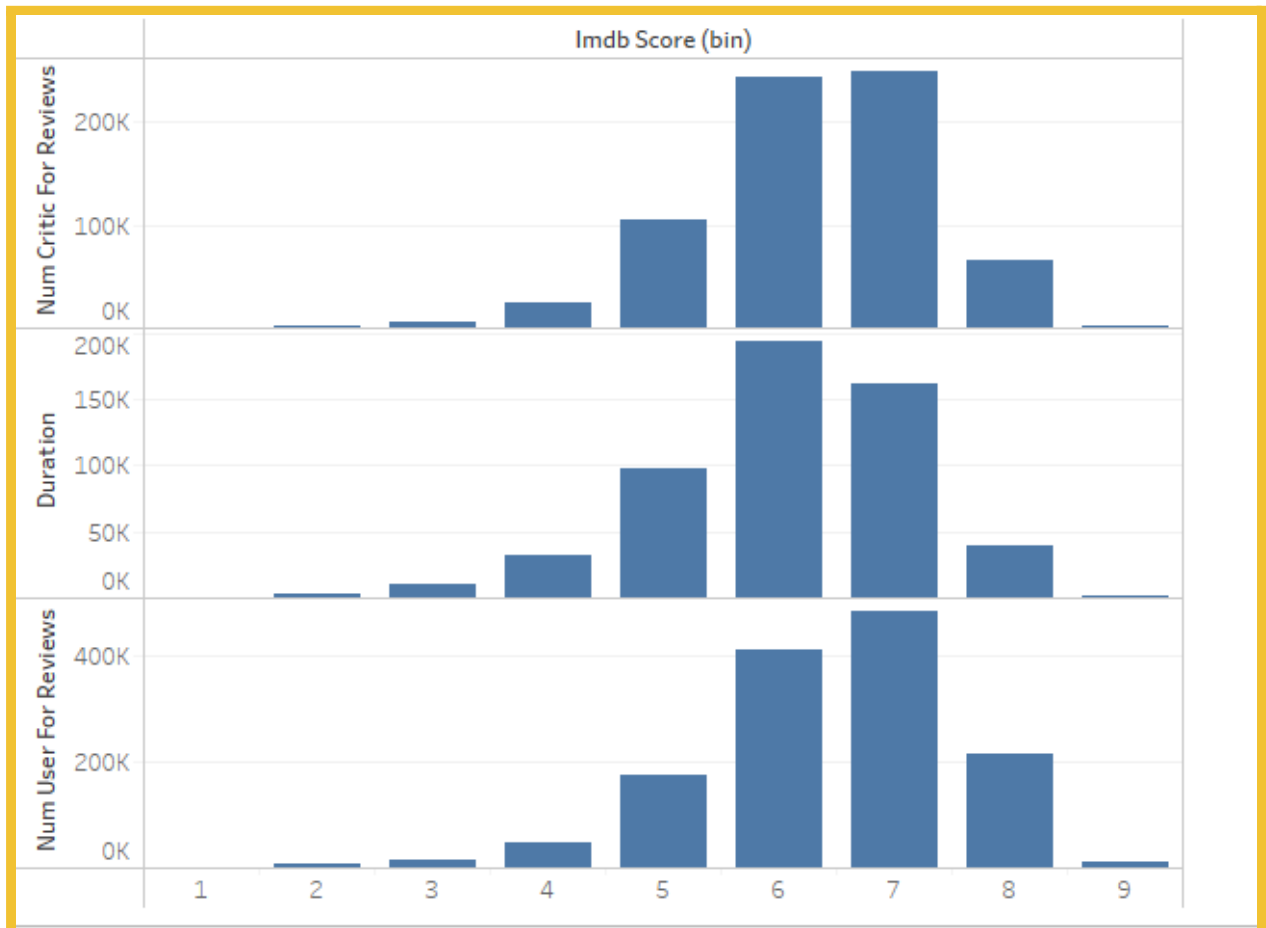
- Joon-ho-Bong is one of the finest directors with movie budgets at its peak. He stands top in the list of 5 famous directors of all time.



- People who really like or dislike a particular movie or casting are more likely to express their strong opinions. By default IMDb users see titles listed in the language used for release in their country.



- Critics - reviews, user reviews vs IMDb score are all interdependent. It interprets and evaluates a film and its place in cinema history, comparing with other films.



THANK YOU!