A CASE STUDY

OF

MOVIE DATA SET CLASSIFICATION

TEAM

AASHITHA	19BCD7159
HRUSHIKESH CHOWDARY	19BCI7073
SAI RAVI TEJA	19BCE7150
VISHNU BHARADWAJ	19BCE7478

Introduction

People’s love for movies is nothing new, be it boredom, or when one is feeling stressful or if one feels like hanging out with friends, movies have become the go-to entertainment. Watching a good documentary or a thought-provoking movie can help you boost your emotional intelligence.They also make us think about certain issues like social injustice, poverty, political power games that we do not often think about.

That is why it is very essential for movie makers to choose the right genre, one that not only triggers testosterone, empathy and other emotions like hope, aspirations, or even fears but also one that has high box office sales.

But the thing with movies is they are not one one dimensional at all,it means a movie doesn’t always have only one genre but can span several genres meaning it's multi dimensional. And in this project, we deal with a large movie data set preferably from IMDB which makes the data set available for free for all the enthusiasts out there and identify the genre that’s highly watched to date.

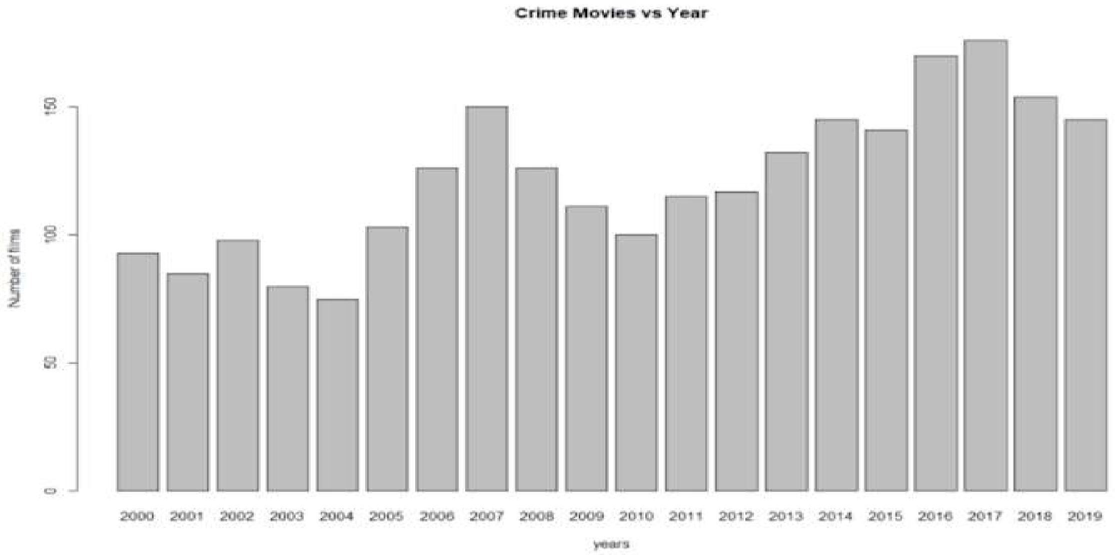
Body

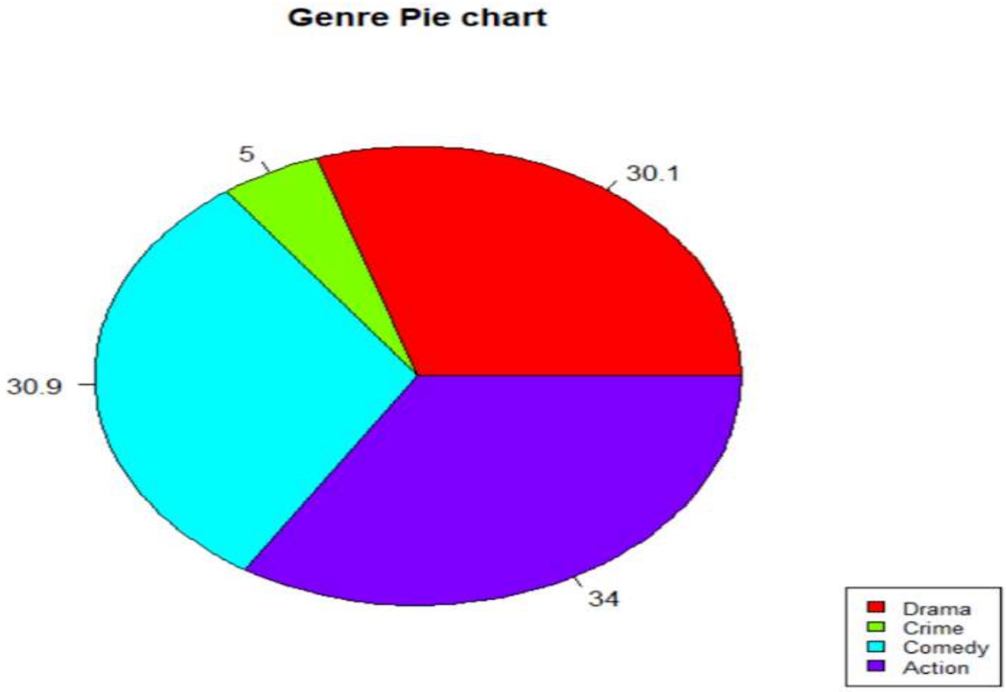
We take an IMDB dataset having more than 50K movie reviews. We use this IMDB movie data set as it contains more precise and accurate data than any other datasets available online. We clean, classify and after analysing this data set, we identify the genre with the highest success rate. This data set has information of the Movie title, original title, year, director, votes, average vote, reviews from the critics as well as from the users, language, actor, genres, episodes, production company, and release date. Missing Values, Outside of key fields, missing values are common. Sometimes the data seems to be unavailable, sometimes it hasn't been entered. Some information is inherently incomplete. Censored Data is ignored.

This is a humongous data set with 22 columns and 85,855 columns and before we start analysing, we must do data cleaning. We must remove the duplicated data, fill in the missing values, and hide the sensitive data from the user and as mentioned above movies are multi dimensional which means we have to convert the multi valued attributes like genre to single valued attribute so that the analysing job is done easier.

Then, we enrich the data which is updating the value entries for deeper understanding and then, we visualize the data we cleaned and pre-processed in the form of bar graphs, pie-charts and line charts so that the reader may readily recognize patterns or trends. From these charts, we can see which groups are highest or most common, and how other groups compare against the others.

From the below bar chart, we can deduce that crime movies are mostly watched in 2017.





The pie chart above shows that Indians preferred action films and least watched the crime films since 2010.

All graphs mentioned above are made by our team to visualise the movie dataset in R-Studio

From the above pie chart, we can say that producing a crime-thriller might result in a commercial failure in India, so action movies are a better option as they have many viewers.

Conclusion

In our project, we represented the genres most watched across the world in the form of graphs for various countries across the world after intensely cleaning, pre-processing and analyzing the data using R-programming so that users can understand it easily by watching it. When a movie is considered a flop and loses money which most movies do by the way, the producers and financiers are the most affected, so one can use this movie data set analyzation and generally invest in movies with genres that have a high success rate.

Appendixes

These links below lead to websites that may not be related to our project but helped us understand in depth about our project

https://arxiv.org/ftp/arxiv/papers/1209/1209.6070.pdf

https://www.analyticsvidhya.com/blog/2019/04/build-first-multi-label-image-classification-model-python/?utm_source=blog&utm_medium=predicting-movie-genres-nlp-multi-label-classification