TERM PROJECT-IMDB DATASET

1) IMDB DATASET

The dataset consists of the following 28 features:

"movie_title" "color" "num_critic_for_reviews" "movie facebook_likes" "duration" "director_name" "director_facebook_likes" "actor_3_name" "actor_3_facebook_likes" "actor_2_name" "actor_2_facebook_likes" "actor_1_name" "actor_1_facebook_likes" "gross" "genres" "num_voted_users" "cast_total_facebook_likes" "facenumber_in_poster" "plot_keywords" "movie_imdb_link" "num_user_for_reviews" "language" "country" "content_rating" "budget" "title_year" "imdb score" "aspect ratio"

It contains data of around 5000 movies spanning across 100 years in 66 countries.

2) ACTIVITY -2 – STORY TELLING

1) Title Year vs Budget – Budget increased over a period of time
2) Content Rating vs Facebook likes – Majority of the Movie going audience go to the PG-13, PG and R rated movies. The plot suggests the same.
3) Country vs Number of Movies – USA is the biggest producer of Movies and this is evident from the plot.
4) Movie distribution in each language – Most of the movies are in English.
5) Gross vs Facebook likes – Generally, highest grossing movies tend to have high social media following.

By doing the above data analyses, we realized that the results were obtained as expected in case.

3) ACTIVITY -3 – DATA PRODUCT IN RSHINY

We implemented a data product in Rshiny as shown below:

In the first tab, a user can have quick look at the statistics of the movies dataset such as

1) **Top 10 Movies** by different selection criteria – Gross, Movie Facebook Likes, Imdb score, Budget and Cast Total Facebook Likes. A user can also look at the results in various movie eras by varying Year Range.

2) We implemented the **Map Reduce Algorithm** on the columns genre (which is of the form as shown below in the dataset) and obtained the total gross collected over the years for each genre. This is done to infer the highest N grossing movie genres.

3) The relation between movie **facebook likes and gross** – Upon observing the graph , it can be inferred that movies made after the year 2000 have more facebook likes when compared to the movies made in the last century as expected. The impact of social media following on the gross

can also be observed. The highest grossing movies made in the recent times have a huge fan following in the form of Facebook Likes. The results can also be seen in different categories of content ratings.

In the second tab, we implemented the **Linear regression model** .

1) Simple Linear model has been implemented to predict the gross based on budget. The predicted point has been marked on the plot

2) The variability of the linear model based on the size of the dataset has been depicted.

Link to App: https://moviesdataset.shinyapps.io/imdb_rating/

Authors:

1) ARUN CHANDRA PENDYALA
2) ABHILASH VELALAM