Good afternoon everyone, my name is nidhi and this is my teammate charan. Our project is on IMDB movie ratings dataset.

INTRO

•Based on the massive movie information, it would be interesting to understand what are the important factors that make a movie more successful than others.
•The combination of user ratings for movies and detailed movie metadata have always been fun to play with.

- As a result, we'd like to look at what types of films are more successful, or have a higher IMDB score. We also want to visualize the outcomes of this study to make them more intuitive.
- In this research, we use IMDB scores as the response variable and analyze the rest of the factors in the IMDB movie dataset to focus on operating predictions.

Statement of goals:

Through a series of plots and inferences, we want to investigate these questions:

1.Relationship between the length of a movie(Runtime) and it's rating on IMDB.

2.Relationship between the count of votes and ratings.

3.Relationship between rating and movies across all years.

4.Are the highest-rated movies the ones with the most votes?

Data Description:

This dataset contains 14 columns for 8451 shows from various countries that spanned 89 years between 1916 and 2005. When we first looked at the Dataset, we noticed that only 4 of the columns contained numerical values, while the rest were categorical, meaning that the data for those columns was stored in a labelled manner.

We deleted outliers (97 quartile runtime values) that potentially skew the distribution in this dataset by removing redundant information such as observations with numerical null values. For the sake of this study, we solely looked at movies.

These are the categorical variables:

Coming to the numerical variables:

Rating – On internet movie database, all films are given an overall rating out of ten.

Vote - All registered IMDb users can submit a single rating – a number between one and ten – for any film on the website. These votes are then rearranged so that certain demographics (newly-registered users, for example) don't disproportionately influence the overall ranking of the film. IMDb ratings are based on the votes of the website's users.

EDA:

Slide 6

Firstly, we want to Investigate from simple histograms of movies Vs year, rating Vs movies, runtime Vs movies, votes Vs movies:

This plot depicts the relationship between count of movies across all years:

- The number of movies released is less during the great depression period(between 1931 to 1939) as compared the movies released before and after the period whereas the World War 2(September 1, 1939 – September 2, 1945) didn't have much affect on the shows count.
- We can also infer that the number of movies being released are exponentially increasing over time.

Slide 7:

This plot describes the the ratings of the movies:

•This distribution is left-skewed because it attains a peak towards the right side and has a tail towards the left side. And also, the mean reflects the skewing the most. Generally, if the distribution of data is skewed to the left, the mean is less than the median which can be clearly seen in the graph.
•The peak is around 1100 meaning that most of the movies have rating of 7.
• and we can also observe that 70% of the movies have the rating between 6 to 8.

The rest of the analysis will be explained by charan