Team 11: Siddharth Sundar, Rochita Das, Kareem Hirani

Introduction and Motivation

Media entertainment such as movies and tv shows has been a popular activity for people for the past couple of decades. Our team wanted to find interesting trends in this domain and we were able to find a variety of datasets regarding movies and tv shows available on popular streaming services and platforms. These datasets contain interesting attributes such as movie names, casts, ratings, countries where they're shown, and much more which will provide us with many opportunities for analysis. Our hope was to create interesting data visualizations and develop key insights that incorporate information from our datasets such as gleaning how some of these attributes relate to another, and what preconceived ideas we have regarding the success of movies are true or untrue. The datasets that we used for our analysis include the following:

1. TMDB 5000 Movie Dataset : This dataset contains information on 5000 movies with no specific time period. It has 13 columns split over two csv files, credits and movies.

   Columns:
   - Title : Movie Title
   - Cast : JSON of actors and actresses
   - Crew: Json of crew
   - Budget: Total cost of development
   - Genres: Genres associated with film
   - Homepage: Link to movie promotional page
   - Movie ID: Unique movie ID
   - Keywords: Relevant keywords regarding themes of movie
   - Original Language: Language in which movie was released
   - Original Title: Original movie title
   - Overview: Brief synopsis of movie plot and premise
   - Popularity: Numerical value of relative popularity of movie
   - Production Company: The production company who produces the movie
   - Release Date: First official release of the movie
   - Revenue: Total revenue (in Dollar)
   - Runtime: Total duration of the movie (in Minute)
   - Spoken languages: List of different languages used in the movie
   - Tagline: A short description to capture the essence of the movie
   - Vote Average: Weighted average of votes
   - Vote count: Total number of users who voted

2. Netflix Movies and TV Shows : This dataset contains information about various movies and tv shows streamed on the Netflix platform. The dataset contains 12 columns and it is all in one file.

   Columns:

   - show_id : Unique ID for every Movie / Tv show
   - type: identifier - a Movie or TV show
   - title: Title of the Movie / Tv Show
   - director: Director of the Movie
   - cast: Actors involved in the movie / show
   - country: Country where the movie / show was produced
   - date_added: Date it was added on Netflix
   - release_year: Actual Release year of the movie / show
   - rating: TV rating of the movie / show
   - Duration: Total length of the movie (in Minute)
   - Listed_in: A movie is classified into different genres
   - Description: Brief summary of the movie

3. Amazon Prime Video Movies and TV Shows
4. Disney+ Movies and TV Shows
5. Hulu Movies and TV Shows
6. https://www.the-numbers.com/movie/budgets (Movie Budget, Revenue info are webscrapped) and Profit is calculated.

Datasets 3-5 have the same columns with respect to Dataset 2. The only difference is the streaming platform.

Visualization Design (Implementation)

The Choropleth Map was created by parsing the combined datasets we had collected, and involved merging incompatible columns, specifically country names. Varying datasets used different names for the same country, for example, “United States” and “United States of America” or countries that on longer exist, for example, “East Germany" and “Soviet Union”. Therefore there was a need to normalize these columns for an ease of analysis and thus was able to easily count the number of movies per country. However, we had the issue that the countries names were not compatible with the choropleth map, since it used the three letter tags instead. This was accomplished through leveraging the python library ‘pycountry’ which allowed the further translation of country names into their three letter tags, which was how our choropleth program identified countries.

Top Directors and Actors are shown using the Bar Plots based on user defined variables (like No of movies, revenue, profit, movie rating etc.) For a better understanding, information is produced as table format as well. We further explored the different genres associated with these top Actors and Directors, Circular Bar Plot is produced to show the counts of different genres.

We have considered movie recommendations as predictive modeling. The model offers generalized recommendations to every user, based on movie popularity and/or genre, actor, director. The basic idea behind this system is that movies that are more popular and critically acclaimed will have a higher probability of being liked by the audience. Here we have calculated weighted movie ratings based on voting average and vote count.

We have also explored the distribution of different continuous variables and how they relate with other other variables as well.

Methodology

Our heatmap can be used to answer the research question "What countries tend to produce more movies within a given year?" The heatmap has dark shades to indicate more production of movies and light shades to indicate less production of movies. This easily helps the user answer the question mentioned earlier.

We have used Bar plots to have a clear visual of top k (user input) Directors and actors. Here we worked on the research question like, " In Depth analysis of top K Directors based on a specific variable say, total revenue from the movies he/she has produced". Also, we have explored if the Director has any inclination to a specific genre. We have created a Circular Bar plot to explore it.

The movie recommendation engine can easily be used by the user to figure out what movies they would like to watch based on a movie that they had already seen and the number of movies they want to see in the results. The results table has the recommendations in descending order of similarity score which helps the user clearly see the K recommended movies that he or she may potentially want to watch.

We used a density plot, QQ Plot, Summary Table, and Scatter Plot (with correlation) Matrix primarily for data scientists to see the kinds of relationships amongst the various attributes in the data. The specific attributes we want the data scientists to look at are Revenue, Budget, Profit, and Popularity Index.

Evaluation Plan

In order to evaluate our user interface, we feel it would be appropriate to conduct user studies. The target audience for the user study would be anyone who has an interest in data visualization and analysis. As indicated by the name, these studies would be individual between one of us and the participant and would take approximately one hour to complete. The participant would be given a google form containing various questions that he or she would have to answer by using our dashboard. As the researcher, we do not want to influence their responses, so we will try to minimize our attempts for assistance unless it is absolutely necessary. At the same time, we will record their computer screen in order for us to see how they interact with our dashboard. After the data analytics questions are answered, we will gather some feedback regarding what the participant thought about the user interface. This will be in the form of an informal conversation and manually recording these responses. We want to gather as much insight as possible to help us see if our dashboard meets the expectations that we intended it to have.

Discussions & Future Work

Going forward, our project could be improved and built upon by designing a more robust recommender system which can take into account more user-facing characteristics such as favorite actor or favorite genre. Currently, it only takes as input the movie that the user likes, and based on that input, a similarity score and weighted rating is computed across all movies in our datasets. The weighted rating is calculated based on the voting average and the vote count which is a decent metric as a basis but it may not be enough with respect to a recommender system. We feel that the layout of our user interface could be improved. We do have navigation tabs which split up our visualizations to an extent, but the pages themselves are very dense in nature. This could be very hard for website visitors to understand each and every visualization as they could be overwhelmed. An improvement would be to have sub tabs within the general navigation tabs as that could break up the visualizations into multiple pages.

References

- **Shiny App References:**
  - https://towardsdatascience.com/end-to-end-dashboard-in-r-shiny-app-64c40d0351d8
  - https://rstudio.github.io/shinydashboard/appearance.html

- **Analysis Refernces:**
  - https://www.kaggle.com/erikbruin/movie-recommendation-systems-for-tmdb

Task Breakdown

Below we have listed tasks that were done by each team member.

Rochita Das

- Data Parsing, Data Cleanup, and Data Merging
- Data Analysis
- Movie Recommender Visualization
- Visualization for both Actor and Directors
- Continuous variable visualization
- Shiny Dashboard Integration

Kareem Hirani

- Data Cleanup (filling in revenues and budgets that were previously NA in dataset)
- Data Analysis
- Choropleth Map Visualization (for spatial component)
- Shiny Dashboard Integration

Siddharth Sundar

- Finding Data Sources
- Data Parsing (formatting dates)
- Choropleth Map Visualization (for spatial component)
- Shiny Dashboard Design