Link to dataset: https://www.kaggle.com/datasets/willianoliveiragibin/10000-data-about-movies-1915-2023/

The IMDb dataset used in the project contains the top 10000 movies released between the years 1915 and 2023. The data contains comprehensive information about the movies such as the name, year, user ratings, metascore ratings, gross income, votes, runtime, genres, etc. I found this data interesting as it could be used to observe trends and the evolution of the film industry over a century and help viewers to find recommendations for new movies based on the various parameters provided. Using this set, I created a system that provides movie recommendations using k-means clustering for the user. The system uses the function to obtain the user's preferred genre and following this, a function for k-means clustering is implemented where clusters are formed based on movie genres and ratings. A function is called to display the recommended movies after a process of filtering and highlighting the common clusters based on the input criteria, following which the effectiveness of the recommendation system is assessed based on its accuracy of whether the filtered movies fall under the user's genre input.

The choice of k-value can significantly have an impact on the interpretability and quality of clustering results. As the k-value for the clusters increases, the number of clusters would increase so as to provide more specialized and niche recommendations for each cluster. However, while a higher k-value can lead to more detailed segmentation, it can further lead to smaller, less meaningful clusters.

The program should be executed and run, after which the user will be asked to enter the genre of their preference, following which the recommendations for a movie would be provided. There would be a series of descriptive statistics provided for each film, such as user rating mean, metascore rating means, gross income means and votes mean. After this the cluster assignment would be given and the recommendations would follow in the order of movie title, genre, run time and the certificate rating which is indicative of whether a movie is rated PG-13 or R to determine how family-friendly it is.