Final Project: Proposal/Dataset Selection

Morgyn Joubert, Sean McLean

College of Professional Studies, Northeastern University

ALY6015: Intermediate Analytics

Prof. Thomas Goulding

January 19, 2024

For film analytics, understanding user preferences and the financial dynamics of movie production is paramount for making informed decisions. This essay explores two key questions using a comprehensive movie dataset: (1) Is there a correlation between the user interactions and the movie features in the dataset that would be crucial for recommendation to a user? and (2) Is there a significant correlation between a movie's budget and its popularity, and can this relationship be used to make informed predictions or recommendations for movie budgeting strategies?

The dataset, "Movie dataset," shows diverse movie attributes such as genres, budget, user interactions, and original language. The focus is on discerning patterns that could shape the film industry's strategies.

Question 1: Correlation between User Interactions and Movie Features

To address the first question, an initial step involves data exploration. Descriptive statistics will unveil the distribution of user interactions (ratings, views, comments) and movie features (genres, language, keywords). A correlation analysis employing the Pearson coefficient will then quantify the relationship between user interactions and various movie features. Visualization tools, such as scatter plots and heatmaps, will be instrumental in providing a visual representation of the discovered correlations.

Question 2: Correlation between Movie Budget and Popularity

Turning attention to the second question, the exploration begins with a detailed analysis of the distribution of movie budgets and popularity scores. A correlation analysis, again utilizing the Pearson coefficient, will be performed to uncover any significant linear relationship between movie budget and popularity. Further, a regression analysis will be employed to model this relationship, offering predictive insights into how budgets may impact a movie's popularity. Visualizations, including scatter plots with regression lines, will be created to enhance the clarity of the correlation.

Both questions require the use of R code to implement the methodologies outlined above. The final project will incorporate R code and outputs, ensuring transparency and replicability. The analysis will not only present results but also offer a detailed interpretation, focusing on the significance of the findings. Visualizations will be carefully chosen to illustrate patterns and trends, making the analysis accessible to a broader audience.

In conclusion, the chosen movie dataset presents an exciting opportunity to delve into the intricacies of user preferences and financial dynamics in the film industry. By employing robust statistical methods, the analysis aims to unearth correlations that can inform decision-making processes, providing valuable insights for recommendations and strategic planning in the ever-evolving world of cinema.

References

Bibin, T. B. (n.d.). **Movie Datas**et. Kaggle. Retrieved from https://www.kaggle.com/datasets/bibintb/movie-dataset?select=movie_dataset.csv

Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.

Kabacoff, R. I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd ed.). Manning Publications.

Northeastern University – Canvas – (Panopto) videos by Prof. Thomas Goulding