Andrew Vick
James Lee
Jessica Price
Samrawit Basazinew

Project 1 Proposal

1. Dataset
    a. https://www.kaggle.com/datasets/lukelarue/movie-attributes-for-3400-movies-from-20002020
    b. For our group project, we have decided to analyze box office earnings for the theatrical run of movies. We feel that it is a subject matter with an abundance of available data as well as multiple factors to analyze. Our primary focus will be on movies from 2001-2021. While box office earnings seem pretty straight forward, we would like to explore the intricate relationships movies have with geography, age restriction of consumers, and production budgets.
    c. https://www.boxofficemojo.com
2. Questions
    a. What kinds of movies do well in the United States versus the global box office?
    b. Do different age ratings (G, PG, PG-13, R) affect the box office profits for movies?
    c. Lastly, which movies are truly the box office winners when we take a deeper look into production budget versus box office earnings?

3. Other Code / Articles
    a. https://www.kaggle.com/code/guillaumes/exploratory-data-analysis-3k4-movies-2000-2020
    b. https://deadline.com/2024/01/international-box-office-2023-global-studio-rankings-market-share-1235709538/
    c. https://www.thewrap.com/pg-13-vs-r-movies-how-each-rating-stacks-up-at-the-box-office/
    d. https://www.the-numbers.com/movie/budgets
4. For each of our research questions, we will be using a data frame to list the movies with their corresponding earnings. To analyze domestic vs international box office, we feel that a good way to present that data comparison is with pie charts (Domestic vs International) and grouped bar charts. We can use the bar charts to compare genres' performance for Domestic vs International. We will be using a box plot to show the distribution of box office earnings across different age ratings. Finally, we will be incorporating a scatter plot to show which movies earned the most once we take their production budget into account.
5. A linear regression model will be used to predict box office success based on its budget. Once we analyze all the data, we should be well quipped enough to make an educated assessment on what movies will see the most success.
6. The pie charts will incorporate gold (US) and teal (International) for its appealing contrast to one another and be the project's main theme colors going forward.

The grouped bar charts will use red (action), teal (drama), and gold (comedy) to represent the different genres. The box plot will use teal for the interior of the box, gold for the median line, and red for the interior of any outliers. The scatter plot and linear regression model will incorporate gold for the data points (movies) and teal for the regression line.

7. Roles & Responsibilities
    a. Andrew – Research, data cleaning, analysis writeup, & presentation slides
    b. James – Question 2
    c. Jessica – Question 1
    d. Samrawit – Question 3 & Linear Regression Model
8. https://github.com/SirBajaBlast/project-1-group-18