Project 2 Proposal: Movie Success Predictors

Team Members: Ethan Duncan, Andy Fiegleman, Arisa Nguyen

GitHub Repository: Project2_Duncan_Fiegleman_Nguyen

Dataset: IMDB Movies with Over 100 Votes (Updated 1/1/2020)

Overview:

We are interested in exploring the IMDB movies dataset to understand the main predictors of movie success. The dataset we will be using scrapes IMDB (a popular movie ratings website) for all movies with over 100 votes or ratings as of 1/1/2020. This dataset contains nearly 86,000 movies with information about each movie. It contains the following information:

- movie's title
- original title
- year of release
- date published
- Genre
- Duration
- Country
- Language
- Director
- Writing
- Producer
- Actors
- Description
- average IMDB rating
- Votes
- Budget
- USA gross income
- worldwide gross income
- Metascore
- number of reviews from users
- number of reviews from critic

This dataset is constructed using four separate datasets shown in the link above (IMDB Movies, IMDB Names, IMDB Ratings, IMDB Title Principles). These tables will be joined on the field imdb_title_id, the names dataset will likely be excluded unless we later find it pertinent to our analysis.

Exploration:

For this project, we are interested in how different aspects of a movie affect that movie's success. We define success as gross profit (adjusted for inflation) and IMDB rating. We will also be considering the number of ratings a title receives should we find that this field be statistically significant to our analysis. The potential predictors we will be focusing on include budget, genre, and duration of the movie. The budget and gross profit of the movie will be adjusted to today's value using the inflation rate of the dollar since that movie was released.

The dataset will need to be cleaned, because not all movies have a listed budget, genre, or duration. The dataset will also need to be analyzed for duplicates and any duplicates will need to be removed. We will also narrow the focus to movies that have been released in the past 10 years and to only American movies.

Some plots we will have are budget vs. gross profit, budget vs. IMDB rating, genre vs. gross profit, genre vs. IMDB rating, duration vs. gross profit, and duration vs. IMDB rating. We envision most of these plots being scatterplots with a linear regression trendline, but may also do a bar graph if we decide that the count of movies is more insightful than a scatterplot showing the spread of movies across the plot. We will also have a table showing the R-value between each set of values.

Our analysis will help us understand whether genre, budget, and duration has the greatest correlation with IMDB rating and gross profit, or if they even have a correlation at all.

Final Report:

Our final 8-10 page report will cover (1) the question(s) we set out to answer (in this case, the strongest predictors of a movies success), (2) how we went about answering this question (data sources chosen, cleansing process, etc), and (3) our ultimate conclusion. We plan not to just answer whether or not our focus variables (i.e., budget, genre, duration) have an effect, but to measure their effect against each other to determine which variable is most correlated with a title’s success, in relation to profit and IMDB rating. In addition to this, we will provide some analysis on the degree to which profit and IMDB rating themselves are correlated.

As mentioned previously, we will provide multiple charts and/or visualizations as part of our final report to support our findings. These charts/visualizations include, but are not limited to, the examples mentioned in the Exploration section of this proposal as we may expand what we choose to add as a result of our analysis. For example, should we find that the number of votes a title receives is statistically significant to that title’s defined success, we shall add that support verbally and visually to our report.