MSDS597 Data Wrangling Project Proposal - Spring 2021

Shang-Hao Huang (sh1384)

April 2021

1 Broad Goals

This project's broad goal of this project is to explore the IMDB movie dataset and address exciting issues such as the genre of movies that are mainly produced, IMDB score analysis, and profitability analysis. Moreover, I want to identify what set of features contribute to a highly rated/profitable film most significantly and try to predict a movie's profitability, if possible.

2 What To Investigate & Why

Although movies are not produced purely for making money, not making sufficient profit puts filmmakers in an awkward financial situation, making them difficult to keep making high-quality films. Therefore, it is essential to determine what common features are among those profitable movies. Furthermore, filmmakers care about not only if the film is profitable, but if it is well acclaimed among critics, so they must identify critical factors amid those films. As a data scientist who enjoys watching all sorts of movies, I want to help tackle these tasks by leveraging my knowledge in data science.

To achieve the objective, some fundamental questions about the data need to be answered. These questions are divided into three parts:

1. **Genre-wise:**
   - What are the most-produced genres of movies?
   - What genres have higher average budget/revenue/profit?
   - How does genre of a movie relate to its the profitability/IMDB score? Is there any correlation?
2. **Country-wise:**
   - Which countries produce most movies?
   - Which countries have the highest average budget/revenue/profit?

3. **IMDB score:**
   - What is the distribution IMDB score?
   - Is there any correlation between score and budget/revenue/profit of a movie?

3 Data Sources

The dataset used in this project is the IMDB 5000 Movie Dataset from Kaggle, which can be accessed via this link: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset. This dataset recorded information on 5043 movies across 66 countries from 1916 to 2016. The dataset is available in a csv format file and is of size 1MB.

Each record in this dataset has 28 variables, including information such as "Title of the movie", "Name of the movie director", "Country the movie was produced in", "Budget of the movie ($)", "profitability", and "IMDB score for the movie (out of 10)".