**Final Project Deliverable 1**

**Angela Wilburn**

**Information Technolgy, Walsh College**

**IT 445 VI Spring 2025**

**Chris Heiden**

**April 22, 2024**

**Company: Wanted Brothers Studios**

**Problem: Identify the correlation between movie characteristics and user sentiment.**

**Problem description and summary**

Wanted Brothers Studios' marketing team has tasked the data analytics team with investigating the correlation between user sentiment and specific movie characteristics using data from the Internet Movie Database (IMDB). The primary goal is to analyze user reviews and determine if specific movie attributes, such as genre and, move rating, run length, are associated with more positive or negative sentiment. Additionally, the team aims to identify common characteristics shared by movies that receive predominantly negative reviews.

By examining the relationship between sentiment and movie characteristics, the marketing team wishes to answer the following questions: Which genres consistently receive more positive sentiment from audiences? Are there specific genres or combinations of genres that tend to have higher rates of negative reviews? Do movies with certain run times or movie ratings have a higher likelihood of positive or negative sentiment? Are there any notable sentiment trends based on the movie's release year or decade?

**Problem reasoning**

The marketing team is responsible for identifying which movies are marketable and have the potential to succeed in the coming years. Wanted Brothers Studios experienced a significant drop in box office sales due to the COVID-19 pandemic. Consequently, corporate has tasked the marketing team with vetting every movie idea to ensure profitability and maximize revenue.

To achieve this, the marketing team will rely on the data the data analytics team discoveres. This data will create a comprehensive strategy to guide the marketing team in deciding which movie ideas to pursue. The team will evaluate each movie idea based on its potential for profitability, target audience, and market demand.

The ultimate goal is to ensure that Wanted Brothers Studios invests its resources in movie projects that will likely succeed in the market. This strategy will allow the company to maximize its profits while minimizing the risks associated with producing movies that may not be successful. By leveraging data insights, the marketing team will be better equipped to make informed decisions and drive the company's success in the film industry.

**Dataset selection**

The datasets selected for this project will come from two sources. The first source is an IMBDB dataset from IEEE DataPort.org (https://doi.org/10.1109/ICCCS49678.2020.9276893). This data set contains two files: one of 1,700 movie metadata(titles, release date, IMDB rating, run time, and genre) and the other of 931,564 unique reviews for those movies. The second source of the dataset is from Kaggle. (<https://www.kaggle.com/datasets/mwallerphunware/imbd-movie-reviews-for-binary-sentiment-analysis>) This dataset contains 25,000 IMDB movie reviews and a positive or negative rating for user sentiment reviews. < 5 percent sentiment rating is a negative review, and < =7 is a positive review.

**Dataset selection reasoning**

The datasets selected for this project are crucial in addressing the primary goal of investigating the correlation between user sentiment and specific movie characteristics. The first dataset from IEEE DataPort.org contains a comprehensive collection of movie metadata, including titles, release dates, IMDB ratings, run times, and genres. This dataset provides essential movie information, allowing for a thorough analysis of the relationship between sentiment and various movie attributes. By combining this metadata with the corresponding user reviews, the data analytics team can identify patterns and trends in audience sentiment based on specific movie characteristics.

Additionally, the second dataset from Kaggle complements the first dataset by providing a large volume of IMDB movie reviews along with their associated sentiment ratings. This dataset is particularly valuable because it offers a pre-labeled set of reviews, with ratings below 5 percent considered negative and ratings equal to or above 7 percent considered positive. Access to this labeled data enables the data analytics team to train and validate sentiment analysis models more effectively. By leveraging both datasets, the team can comprehensively understand how user sentiment varies across different movies and identify the key characteristics that influence audience reception empowering the marketing team to make data-driven decisions and develop targeted strategies for movie selection and promotion.