**Final Project Deliverable 2**

**Angela Wilburn**

**Information Technology, Walsh College**

**IT 445 VI Spring 2024**

**Chris Heiden**

**May 6, 2024**

**Company: Wanted Brothers Studios**

**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 particular movie attributes, such as genre, move rating, and 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 discovers. 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 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 storage**

Before loading the data into the Final\_Project\_Deliverable\_2 Jupyter Notebook, the combined\_df of movie metadata had duplicate rows of movie titles (see Appendix 1). So, before merging and exporting the reviews with the metadata, the drop duplicates function was used to ensure unique movie metadata (See Appendix 2). After loading the "Movie\_MetaData\_Reviews" dataset, I noticed a few things. The "release\_date" column contained two variables: release date and country. There are missing values in the movie\_rated and rating\_y columns. There are 142 rows of movie data missing the movie rating, and 111,200 of the rows are missing the ratingh\_y or reviewer rating. There are 1,067,384 rows of data and 18 columns. Most of the rows in the data are not valuable for the analysis we will be doing, and some of the columns of the data are ambiguous and need to be changed. The movieReviewTraining dataset is extremely clean, with no missing columns or columns that need to be altered or removed. It has 25,000 rows and two columns.

**Data cleaning/cleansing (removal of unwanted variables)**

The dataset that needs the most cleaning is the Movie\_MetaData\_Reviews dataset. First, I split the "release\_date" into two columns to better analyze the movie metadata features. Next, the dataset is relatively large, so removing the null values from the dataset will not affect the overall analysis. 142 "movie\_ rated" rows and 111,200 "rating\_y" rows were dropped from the dataset. The following columns were removed: the "num\_raters" column because that is not the number of ratings we are analyzing in the data set, and the "num\_reviews" as this number does not match the number of reviews collected for the movies we will be reviewing. The review\_url is unnecessary, and the username will not matter for our analysis; helpful, total, and title are all unnecessary columns. The year column, as we can quickly get that information from the "release\_date."

Once the right columns were selected, the "rating\_x" and "rating\_ y" columns needed to be renamed to less ambiguous names, rating\_x renamed to "IMBD\_rating," and rating\_y being reviewer\_rating. This data cleansing will clean up memory by removing unnecessary columns and allow for better analysis with meaningful names.

With the data now free of null and unwanted columns, the columns need to be changed to the right type to perform statistical analysis. The "reviewer rating was changed to float64 to match the "IMBD\_rating" column.

**Descriptive Statistics**

The data set has changed since the project deliverable one assignment because of the cleaning and manipulation of the data. I still used the same data source, but duplicates and null values have been removed. The descriptive statistics of the "IMDB\_rating" columns show that the mean of IMDB ratings is about 7.5, and the standard deviation for the rating is .746, meaning that most of the ratings fall within the mean. The lowest IMDB score is 3.5, and the highest is 9.3. Conversely, the "reviewer\_raiting" column has a mean of 7.3 but a standard deviation of 2.84; the lowest score is a 1, and the highest is a 10, meaning there is more variability in the individual review ratings. This is valuable information for our problem because seeing if the movie has an overall positive sentiment will allow us to draw insights more quickly than the individual review ratings alone.

The "movieReviewTraining" dataset has only two columns: one with a sentiment label and the other with a comment. It is evenly split with 12,500 positive and 12,500 negative comments and labels. This dataset will help train the machine learning model that will allow us to perform user sentiment analysis on the Movie\_MetaData\_Reviews dataset.

**Data Visualization**

The first of the three charts chosen for this deliverable is the side-by-side histogram chart comparing the distribution of IMDB ratings and reviewer ratings. This chart was selected to illustrate the potential discrepancy between expert opinions and public sentiment when predicting which movies people will watch. The histogram of IMDB ratings demonstrates a normal distribution, which is generally suitable for predicting outcomes. However, compared to the histogram of reviewer ratings, it becomes evident that the two distributions differ significantly. This visual comparison highlights the importance of considering expert opinions and public sentiment when analyzing movie preferences. While IMDB ratings provide a standardized measure of movie quality based on expert evaluations, they may not fully capture the diverse opinions and experiences of the general audience. The reviewer ratings, on the other hand, reflect a more comprehensive range of individual preferences and critiques.

This chart directly relates to our chosen dataset, including IMDB ratings and user reviews. By visualizing the distributions of these two variables, we can better understand the dynamics at play and make more informed decisions when using this data to predict movie performance, including user sentiment.

The second chart chosen for this deliverable is a multiple-line time series chart displaying the number of movies released by year and country. This graph showcases the temporal and geographical distribution of movies in the dataset, allowing us to identify trends and patterns in movie sentiment over time across different countries. The chart enables us to compare the volume of movies across countries, identifying which countries have a higher or lower sentiment and which years. The chart showcases how the dataset has a diverse range of release countries and periods.

The final chart selected for this deliverable is a powerful combination of a bar graph and a multiple-line chart. It displays the number of movies in each genre and their average IMDB and user ratings. This chart serves a dual purpose: it highlights the potential differences between critic and public opinion and the vast range of genres available for sentiment analysis. The chart directly compares expert and audience perspectives by presenting the average IMDB and user ratings for each genre. This visual representation reinforces the idea that critics' opinions may not always align with the general public's sentiments.

The combination bar graph and multiple line chart effectively demonstrates the relationship between movie genres, their popularity, and the differences that may exist between expert and public opinion. It underscores the importance of considering critical and audience perspectives in our sentiment analysis while showcasing the broad genre landscape available for exploration.

Appendix 1 A screenshot of a computer

Description automatically generated

Appendix 2

A screenshot of a computer

Description automatically generated