Film Data Analysis

Group Members: Yu Wang, Yuhao Sun

For this project, we would like to discover the significant factors that contribute to the general success of a movie. We will be looking at the relationship between the gross box office/revenue and features like movie genre, Oscars nominations/award, reviews, soundtrack ratings, and trailer popularity.

The Data used in this movie will be:

* Oscars data: <https://www.kaggle.com/datasets/unanimad/the-oscar-award>
* Trailer dataset: <https://grouplens.org/datasets/movielens/20m-youtube/>
* Soundtrack reviews: <http://www.movie-wave.net/reviews-by-title/>

The main questions for this project will be:

1. What type of movies tends to be more successful at the box office?
2. Is there any correlation between the reviews for the music in the movie and the movie’s success?
3. Does a high movie budget help in terms of the box office?
4. How can we us machine learning to predict the success of a movie? (Stretch)

We will be using data shaping and data frame operations on the Oscars data analysis. We will also be using web scraping for the trailer and music data. We will be doing data cleansing and merging for the gross revenue analysis. Tools like seaborn, matplotlib, etc. will be utilized for the data visualization. Finally, we will be using scikit-learn and Random Forrest Classifier for the machine learning modeling training to predict the success of a specific movie.

We plan to satisfy the following requirements:

* Data is collected through a means more sophisticated than downloading (e.g. scraping, API).
* At least one of the datasets contains more than 1,000,000 rows.
* It combines data collected from 3 or more different sources.
* The analysis of the data is reasonably complex, involving multiple steps (geospatial joins/operations, data shaping, data frame operations, etc.
* You perform a machine learning analysis with scikit-learn as part of the analysis.

**Describing the Project**

# Section 1: Discover Which Film Genre Is Most Likely To Receive The Honor?

Graphical user interface, text, application

Description automatically generated A picture containing text

Description automatically generated

We first reshape the dataframe by trim the data to only genres and title columns, then merge the current dataframe with the Oscars data we got from online. The visual representations were made by seaborn package, and we used scatter plots and bar charts to describe our findings.

Section 2: How does trailer popularity affect movie success

Graphical user interface, text, application, email

Description automatically generated

First we save the scraped youtube data as a csv file, then we combined the data from movie wave and scraped YouTube data as a big dataframe. To get the desired data, we sort the highest ROI in the dataframe and get rid of the duplicates before making the production budget vs return on investment plot.

Section 3: Do music scores contribute to movie success

Graphical user interface, text, application

Description automatically generated with medium confidence Graphical user interface, text, email

Description automatically generated

## We did data wrangling and cleaning on the raw scraped music reviews data from movie wave website. Then Compile for Composer and Dataframe Stack Genre DataFrame and Merge with Financial Data. Once we have everything we need, we created new columns: year, domestic gross, production budget, worldwide gross, ROI in our dataframe.

## Section 4:

## Graphical user interface, text Description automatically generated

## By using BeautifulSoup, we scraped actor, director data from imdb website first, then combined these data as a dataset for training and testing.

## Graphical user interface, text, application Description automatically generated

The model contains categorical and non categorical variables, we used label encoder and one-hot encoder from sklearn package to implement these different types of variables into our model.

Then developed 3 different methods for machine learning (A Lasso regression, SGD classifier, Random Forest Classifier). We chose the random forest method to be the final model because Random Forest gives us better F1 score.

Text

Description automatically generated

To test our model, we chose the new avatar movie to be the subject, the model requires information such as rating, genre, country, star1, star2, director, runtime, released day in order to work. The result we got from the model showed that this model is going to be a huge success.