Project Proposal

Team name - Team member 1, Team member 2, Team member 3, Team member 4

```
library(tidyverse)
library(tidymodels)
library(knitr)
# add other packages as needed

# add code to load data
IMDB_Movies <- read_csv("data/final_movie_data.csv")

glimpse(IMDB_Movies)
```

```
Rows: 177,022
Columns: 24
$ id                 <dbl> 299534, 475557, 634649, 299537, 429617, 466272, 5~
$ title              <chr> "Avengers: Endgame", "Joker", "Spider-Man: No Way~
$ vote_average       <dbl> 8.263, 8.168, 7.990, 6.843, 7.447, 7.437, 7.990, ~
$ vote_count         <dbl> 23857, 23425, 18299, 14657, 14495, 12237, 11395, ~
$ status             <chr> "Released", "Released", "Released", "Released", "~
$ release_date       <date> 2019-04-24, 2019-10-01, 2021-12-15, 2019-03-06, ~
$ revenue            <dbl> 2800000000, 1074458282, 1921847111, 1131416446, 1~
$ runtime            <dbl> 181, 122, 148, 124, 129, 162, 119, 131, 192, 131,~
$ adult              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ backdrop_path      <chr> "/7RyHsO4yDXtBv1zUU3mTpHeQ0d5.jpg", "/hO7KbdvGOtD~
$ budget             <dbl> 3.56e+08, 5.50e+07, 2.00e+08, 1.52e+08, 1.60e+08,~
$ homepage           <chr> "https://www.marvel.com/movies/avengers-endgame",~
$ imdb_id            <chr> "tt4154796", "tt7286456", "tt10872600", "tt415466~
$ original_language  <chr> "en", "en", "en", "en", "en", "en", "en", "en", "~
$ original_title     <chr> "Avengers: Endgame", "Joker", "Spider-Man: No Way~
$ overview           <chr> "After the devastating events of Avengers: Infini~
$ popularity         <dbl> 91.756, 54.522, 186.065, 50.399, 49.913, 47.706, ~
$ poster_path        <chr> "/or06FN3Dka5tukK1e9sl16pB3iy.jpg", "/udDclJoHjfj~
$ tagline            <chr> "Avenge the fallen.", "Put on a happy face.", "Th~
```

```
$ genres               <chr> "Adventure, Science Fiction, Action", "Crime, Thr~
$ production_companies <chr> "Marvel Studios", "Warner Bros. Pictures, Joint E~
$ production_countries <chr> "United States of America", "Canada, United State~
$ spoken_languages     <chr> "English, Japanese, Xhosa", "English", "English, ~
$ keywords             <chr> "superhero, time travel, space travel, time machi~
```

Introduction

Movies have been at the heart of pop culture for nearly a century. And while movies have changed a lot over the years, varying in length, theme, due to technological advances, and more, one thing has never changed–movies have remained as a dominant part of the entertainment industry. Yet, ever since the COVID-19 pandemic, the North American box office for movies has not recovered–it seems that people just aren't going to movies as much as they used to. Domestic box office gross revenue in 2019 reached nearly $11.4 billion, while the yearly total from 2023 topped off at $8.9 billion after a very slow two years in 2020 and 2021 (Box Office Mojo). With moviegoers back on the rise, the film industry and box offices are hopeful to end the decade on a high note and return back to their original state. As such, we are interested in investigating the trends in gross revenue for top movies and how various elements impact the amount of money that a movie makes.

"Domestic Yearly Box Office." Box Office Mojo, 2 Oct. 2024, www.boxofficemojo.com/year/. Accessed 2 Oct. 2024.

Research Question: What are the key factors and metrics that most significantly influence a movie's gross revenue?

Motivation: As a group of movie enthusiasts, movie popularity has always been a large source of curiosity. We've seen terrible movies whose popularity we find puzzling, and unknown masterpieces that do not get the attention they deserve. What is it that separates the two? Is it primarily the "goodness" of the movie that determines the money it brings in, or are there other factors that are equally influential? Is the 1.4 billion dollar success of Alvin and the Chipmunks (an objectively pretty bad movie) just an outlier, or reflective of a deep and complex relationship between a movie's characteristics and its gross profit? Through our analysis, we hope to get to the bottom of these questions, and gain a deeper understanding of what truly drives a movie's success or failure in the box office.

Hypotheses: We hypothesize movies that have high popularity scores and ratings, substantial budgets, and broad audience appeal will be more likely to generate higher gross revenue. We also hypothesize movies with more than one genre will be more likely to generate higher gross revenue. Movies with a runtime between 1.5-2 hours will be more likely to generate higher gross revenue.

Data description

The source of the dataset is taken from TMDb (The Movie Database), which is an online, public database for movies that contains information such as ratings, runtime, cast, director, and box office performance.

We discovered this dataset at this link on Kaggle: https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies and it is updated daily by user asaniczka, who takes the data from TMDb.

There are 953629 observations in this dataset, each representing a single film. There are 24 columns in this dataset, with some of the most important being the name (an identifier), release year, runtime, genres, rating, votes, and revenue.

Exploratory data analysis

We filtered the dataset because the original dataset was far too large to be able to run efficiently in RStudio: Firstly, we filtered to only include recent movies, from 2019 to 2025. Then, we filtered for movies where the original language is english. Lastly, we filtered for where the movies had more than 100 votes. (EDIT THIS, THIS IS PART OF EDA, AND ADD ON THAT WE WILL FILTER FURTHER BY THINGS SUCH AS RELEASE STATUS)

Analysis approach

...

Data dictionary

The data dictionary can be found https://labs-az-08.oit.duke.edu:30174/files/project-Rated-R/data/README.html