Project Proposal

Rated R - Carlie Scheer, Christina Lee, Jerry Lin, Samir Travers

```r
library(tidyverse)
library(tidymodels)
library(knitr)

library(readr)
movies <- read_csv("data/final_vote_filtered_movie_data.csv")
```

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; as such, we filtered for where the movies had more than 100 votes, meaning more than 100 people submitted ratings for the movie.

We are going to transform the variables spoken_languages and genres. For spoken_languages, we want to make a categorical variable based on whether more than one language is spoken in the movie or not (True = more than one language, False = just one language). For genres, we want to take a similar approach to see whether the movie is tagged for more than one genre (True = more than one genre, False = just one genre). Additionally, we want to transform the month, date, year format of the release date to just be the year. Lastly, we are going to remove all 0s in the revenue column because we are not able to accurately impute the values; we have evidence to suggest that the original author of the dataset just put 0s instead of NAs.

```
movies <- movies |>
  filter(revenue > 0)
  # filter(revenue > 0 & budget > 0) |>
  # filter(revenue > 3 * budget)

movies |>
  ggplot(aes(x = revenue)) +
  geom_histogram(bins = 20) +
  labs(
    title = "Histogram of Revenue",
    x = "Revenue (in billions of dollars)",
    y = "Count"
  ) +
  scale_x_continuous(labels = label_dollar(scale_cut = cut_short_scale()))
```

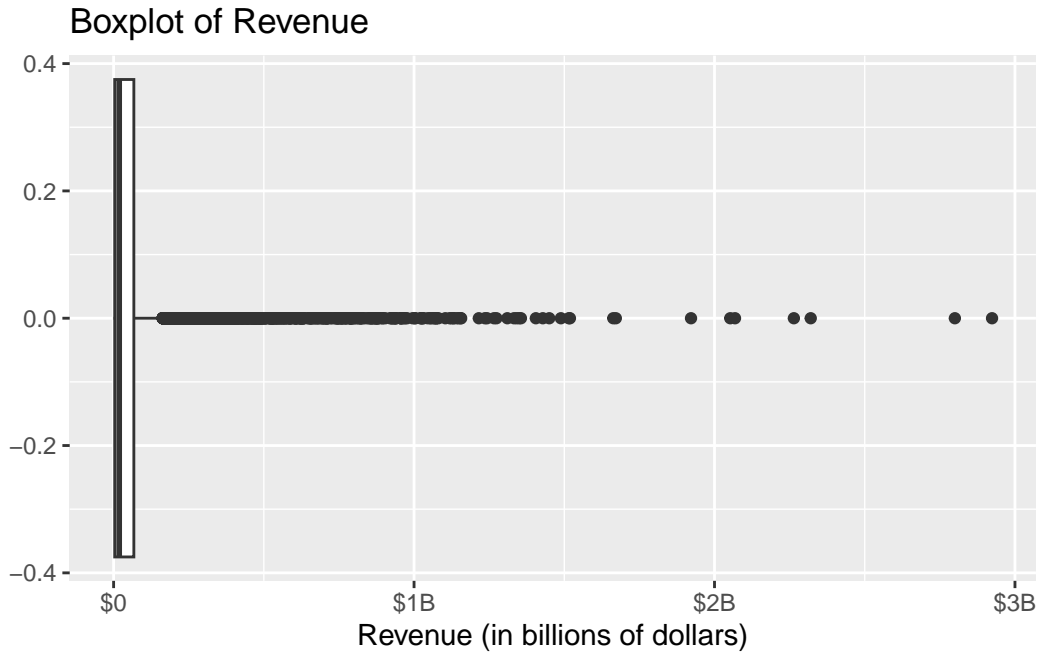Histogram of Revenue
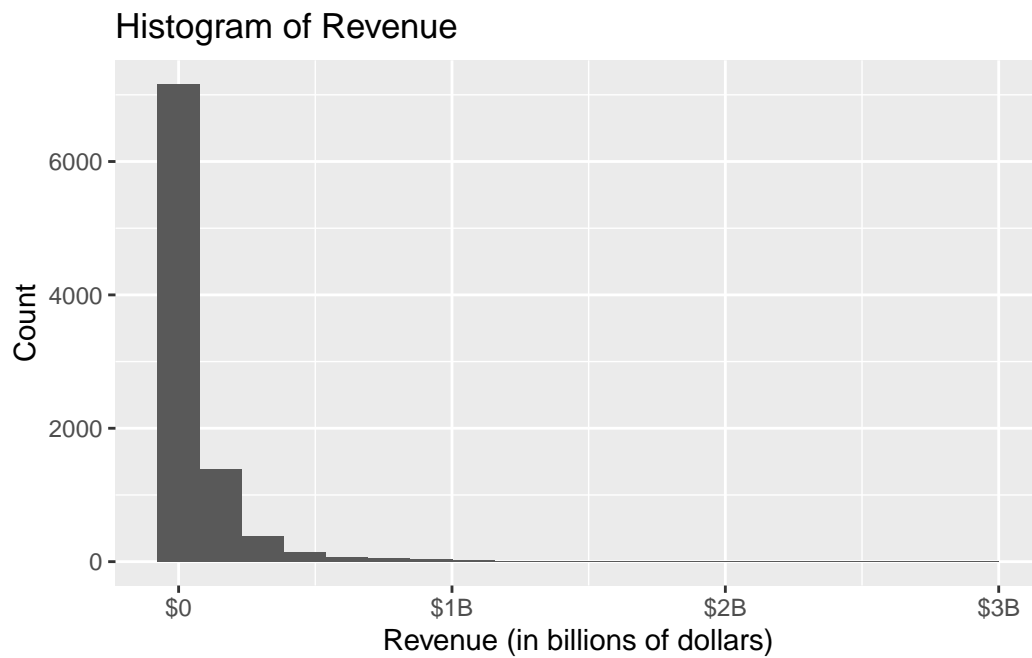
```
movies |>
  ggplot(aes(x = revenue)) +
  geom_boxplot() +
  labs(
    title = "Boxplot of Revenue",
    x = "Revenue (in billions of dollars)"
  ) +
```

```
scale_x_continuous(labels = label_dollar(scale_cut = cut_short_scale()))
```

Boxplot of Revenue


```
movies |>
  summarise(min = min(revenue),
            q1 = quantile(revenue, 0.25),
            median = median(revenue),
            q3 = quantile(revenue, 0.75),
            max = max(revenue),
            mean = mean(revenue),
            sd = sd(revenue)) |>
  kable(digits = 3)
```

min	q1	median	q3	max	mean	sd
1	3667458	18553131	67397992	2923706026	72674288	160447113

The distribution of revenue is strongly skewed to the right. The boxplot suggests that we have a significant amount of outliers. The median is $18,553,131 and our interquartile range is $48,844,861. We will need to adjust our filters in order to get less skewed data.

Analysis approach

Variables of interest:

vote_average: This is the average rating given to the movie by viewers. We expect this to be correlated with the response variable, as good ratings will encourage more people to go see the movie, and if someone likes the movie they will be more likely to recommend it to their friends

spoken_languages: This is the number of different languages spoken in the movie. We hypothesize that if a movie has more languages in it, it will appeal to a broader audience and therefore earn greater revenue.

budget: This is the amount of money spent on producing the movie. We expect movies with a larger budget to earn more revenue, as the production company has more resources to put into the movie.

runtime: This is the length of the movie. We think that people may be more willing to see shorter/average length movies, as it is less of a time commitment for them.

release_date: This indicates the day, month, and year that the movie was released. We expect that the year will greatly impact gross revenue because, for instance, movies released right before or during the pandemic will have very low gross revenue as no one was able to go to the theaters.

genres: This is a genre or list of genres associated with the movie. We expect movies tagged with more than one genre to appeal to a wider audience, thus driving up the number of people who go to see the movie and increasing gross revenue.

Method of approach:

We plan to use multiple linear regression to predict gross revenue.

Data dictionary

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