Project Proposal

The Four-mula: Hellen Han, Lila Rogers, Amy Duan, Eva Aggarwal

```
library(tidyverse)
library(tidymodels)

imdb_data <- read_csv("data/Imdb Movie New Data_v2.csv")
```

Introduction

As a group, we decided to explore several metrics related to IMDb movie revenues to understand what contributes to a movie's overall box office success. We chose revenue as our success metric, as this is the most popular measurement of a movie's success in other academic work, therefore allowing our analysis to complement existing literature. A movie's revenue is defined as the total amount of money a given movie makes from ticket sales, and is influenced by a wealth of factors, several of which we will explore in our analysis. Given our goal of determining what factors contribute most to a movies success, our leading research question is:

What factors in a movie's production and release have the greatest impact on its total revenue?

We chose this research question because in determining what contributes to a movie's revenue, we can gain the necessary understanding to predict how certain movies will do based on various factors. We predict that the factor that will have the greatest influence on total revenue will be budget, as a movie's budget likely contributes to several other influential aspects of a movie such as the actors featured in the movies, the investment in marketing the movie before its release, and the overall quality of the film.

Data description

We obtained this data set from Kaggle, an online data science platform with a collection of community-developed open data sets. Here is the link to where we found the data.

This data was collected by **Anand Shaw** from the **IMDb website** using various IMDb sites, and converted into a .csv file. The data was updated on a daily basis until 2 months ago.

This data set collects information available on the IMDb website for different movies, such that each observation describes characteristics of a specific movie. In general, the characters being measured follow basic information about the movie, various classifications of the movie's popularity and rating, and the monetary values associated with the movie. It measures the following **15 characteristics** per movie:

`title`: The name of the movie.

`vote_average`: The average rating the movie has received from users (on a scale, typically from 0 to 10).

`vote_count`: The total number of votes or ratings submitted for the movie.

`status`: The current state of the movie (e.g., "Released," "Post-Production").

`release_date`: The date when the movie was officially released.

`revenue`: The total earnings the movie made (usually in USD).

`runtime`: The duration of the movie in minutes.

`adult`: Indicates whether the movie is classified as adult content (e.g., "True" or "False").

`budget`: The total cost of producing the movie (usually in USD).

`imdb_id`: The unique identifier for the movie on IMDb (Internet Movie Database).

`original_language`: The language in which the movie was originally produced (e.g., "en" for English).

`popularity`: A metric indicating how popular the movie is (typically based on views, searches, or ratings).

`genres`: The categories or genres the movie belongs to (e.g., Action, Comedy, Drama).

`production_countries`: The countries where the movie was produced.

`spoken_languages`: The languages spoken in the movie.

```
glimpse(imdb_data)
```

```
Rows: 1,048,575
Columns: 15
$ title                <chr> "Inception", "Interstellar", "The Dark Knight", "~
$ vote_average         <dbl> 8.364, 8.417, 8.512, 7.573, 7.710, 7.606, 8.255, ~
$ vote_count           <dbl> 34495, 32571, 30619, 29815, 29166, 28894, 27713, ~
$ status               <chr> "Released", "Released", "Released", "Released", "~
$ release_date         <chr> "7/15/10", "11/5/14", "7/16/08", "12/15/09", "4/2~
$ revenue              <dbl> 825532764, 701729206, 1004558444, 2923706026, 151~
$ runtime              <dbl> 148, 169, 152, 162, 143, 108, 149, 139, 121, 154,~
$ adult                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ budget               <dbl> 1.60e+08, 1.65e+08, 1.85e+08, 2.37e+08, 2.20e+08,~
$ imdb_id              <chr> "tt1375666", "tt0816692", "tt0468569", "tt0499549~
$ original_language    <chr> "en", "en", "en", "en", "en", "en", "en", "en", "~
$ popularity           <dbl> 83.952, 140.241, 130.643, 79.932, 98.082, 72.735,~
$ genres               <chr> "Action, Science Fiction, Adventure", "Adventure,~
$ production_countries <chr> "United Kingdom, United States of America", "Unit~
$ spoken_languages     <chr> "English, French, Japanese, Swahili", "English", ~
```

Due to the large size of the original data set, for the purpose of uploading this to our repository, we dropped 6 characteristics (`id`, `original_title`, `tagline`, `production_companies`, `overview`, `keyword`) from the data set due to their redundancy, irrelevance to the research question and presence of a significant number of null values. Thus, our data set now contains **20 variables**.

Data processing

For the following IMDb movies data set, to prepare the data set for analysis, first, we need to consider missing data values within the dataset, eliminating observations that are incomplete across our predictor/response variables and filtering out movies that haven't yet been released. The majority of these missing values appear in variables that aren't relevant to our regression analysis, however, we still need to account for missing values in our predictor variables.

Furthermore, we need to check for data inconsistencies in the quantitative variables by analyzing the summary statistics for the predictor variables and noting whether there are extreme outliers that are improbable (quantitative variables with an unreasonable value of 0 or negative values). For ease of interpretation, we might scale some quantitative variables, such as revenue and budget scaled in millions of dollars.

In addition, our dataset requires a transformation across a few of the variables to ensure that these predictors can be used in the regression analysis. For example, our categorical predictors, such as adult and original language, need to be converted to factors. Furthermore, the release year needs to be extracted from the release_date variable to create a new quantitative variable that analyzes the number of years the movie has been released since the present year (2025).

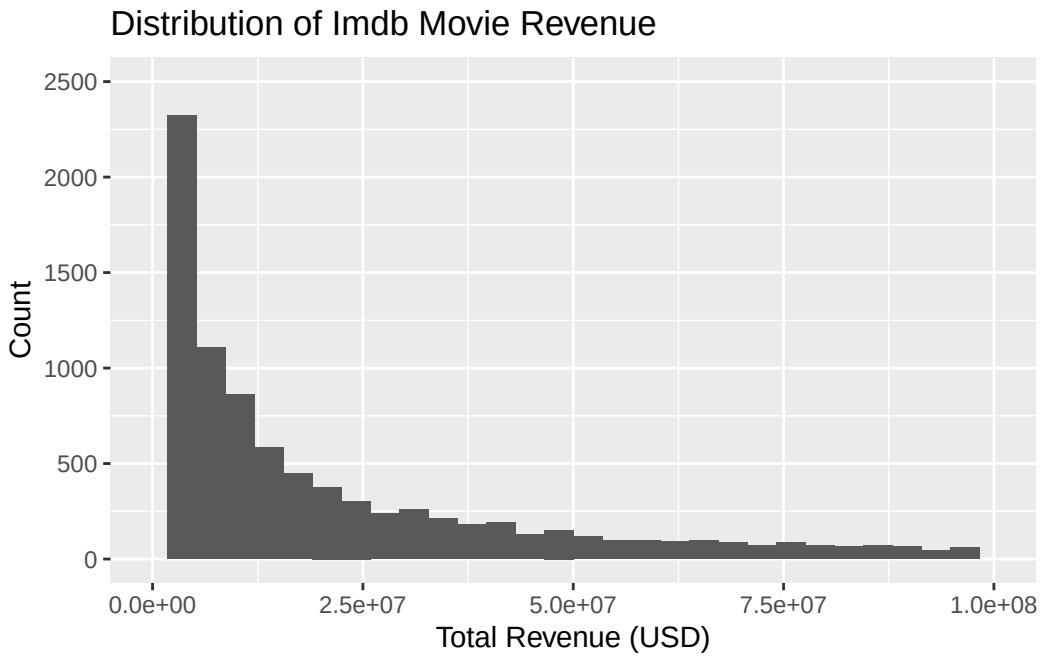
```
summary(imdb_data$revenue)
```

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
       -12          0          0     760411          0 4999999999
```

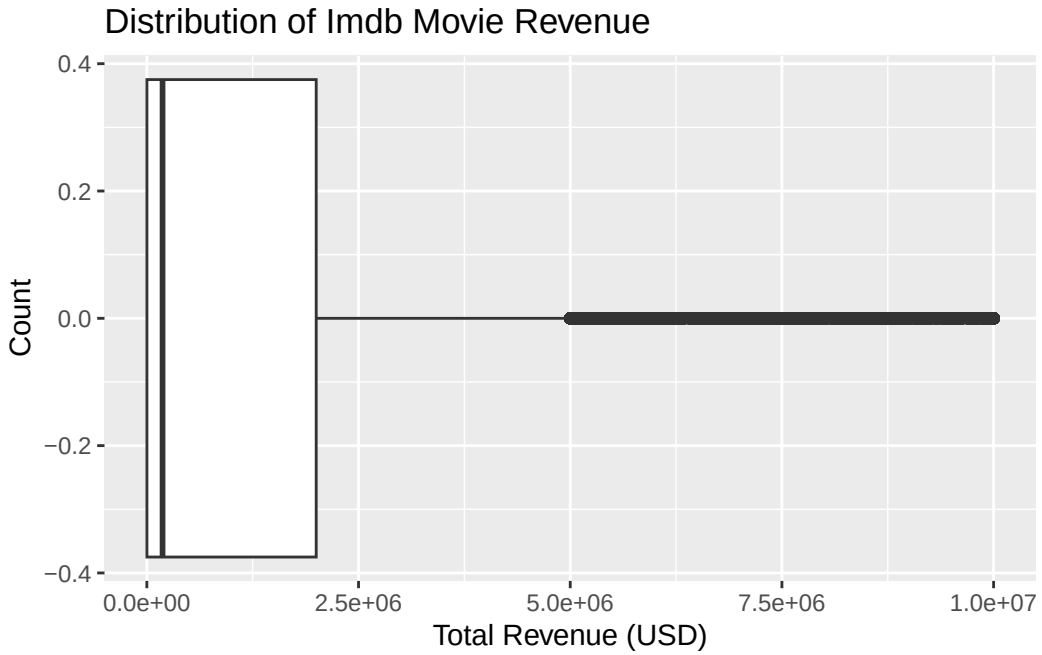
```
imdb_data_revenue <- imdb_data %>%
  filter(revenue > 0)
summary(imdb_data_revenue$revenue)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
1.000e+00 2.400e+04 1.818e+06 3.857e+07 1.864e+07 5.000e+09
```

```
imdb_data %>%
  filter(revenue > 0) %>%
  ggplot(mapping = aes(x = revenue)) +
  geom_histogram() +
  labs(title = "Distribution of Imdb Movie Revenue",
       x = "Total Revenue (USD)",
       y = "Count") +
  scale_x_continuous(limits = c(0, 100000000)) +
  scale_y_continuous(limits = c(0, 2500))
```




```
imdb_data %>%
  filter(revenue > 0) %>%
  ggplot(mapping = aes(x = revenue)) +
  geom_boxplot() +
  labs(title = "Distribution of Imdb Movie Revenue",
       x = "Total Revenue (USD)",
       y = "Count") +
  scale_x_continuous(limits = c(0, 10000000))
```



From the summary statistics, it is observable that the majority of these revenue values are unknown values (represented by 0), as the 1st quartile, median, and 3rd quartile are all 0 dollars. Thus, to obtain representative visualizations of the distribution of the revenue response variable, we analyze IMDb movies that have a revenue greater than 0 (non-missing values in this data set).

The visualizations describe that the distribution of IMDb movie revenue is heavily right-skewed and unimodal, indiciating that there are various outliers in IMDb movies that generate significantly higher revenues. For the purpose of creating meaningful visualizations, we omitted IMDb movie revenues that are greater than 100,000,000 dollars for the histogram and greater than 10,000,000 dollars for the boxplot. Furthermore, the distribution of IMDb movie revenues has a center of approximately 1,818,000 dollars, described by the median, and a spread of approximately 18,616,000 dollars, described by the interquartile range. There are multiple outliers in the dataset for IMDb movie revenue greater than approximately 5,000,000 dollars, as shown in the boxplot visualization, as a few IMDb movies have significantly higher generated revenues. The majority of IMDb movies generate between approximately 0 and 2,000,000 dollars in revenue.

Analysis Approach

Given the information provided by the data set, we are interested in investigating what factors contribute most to the movie revenue (`revenue`) of IMDb movies to provide more insight into recent trends in the movie industry. This could help us understand how producers and directors can maximize their revenue with every movie they work on. Some potential predictor variables of interest include:

- `vote_average`: The average rating the movie has received from users, on a scale from 0-10 can help us understand how popularity within users influence movie revenue
- `status`: The current state of the movie ("Released","Post-Production" etc.) can help us understand whether the state of the movie influence revenue
- `release_date`: The date when the movie was officially released can help us understand the trend of movie revenue throughout the time frame
- `runtime`: The duration of the movie in minutes could also influence movie revenue
- `adult`: whether the movie is classified as adult content can also influence movie revenue, as specific content might attract more audiences than the others
- `budget`: The total cost of producing the movie might indicates how much upfront capital is poured to the production and thereby influence the movie revenue.
- `original_language`: the language in which the movie was originally produced might also influence movie revenue, as specific language might appeal to more audiences and cover a larger group of people than the others and thereby leading to more revenue generation
- `genres`: movie genre might also influence movie revenue. For instance, some genres might appeal to younger and older audiences alike, generating more revenue.

Given that our response variable is quantiative and that we have multiple predictor variables (both quantitative and categorial), we believe a **multiple linear regression model** is the most suitable regression model for answering our research question. This type of model can handle multiple predictors simultaneously and quantify the impact of each predictor on the revenue efficiently.