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 - Lila

- An introduction to the subject matter you're investigating (citing any relevant literature)
- Statement of a well-developed research question.
- The motivation for your research question and why it is important
- Your team's hypotheses regarding the research question
  - This is a narrative about what you think regarding the research question, not formal statistical hypotheses.

Data description - Eva

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 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 is some characteristics of a specific movie. It measures the following **21 characteristics** per movie:

1. `id`: A unique identifier for each movie.

2. `title`: The name of the movie.

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

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

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

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

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

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

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

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

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

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

13. `original_title`: The original title of the movie in its native language.

14. `overview`: A brief summary or description of the movie's plot.

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

16. `tagline`: A short phrase or slogan associated with the movie.

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

18. `production_companies`: The names of the companies involved in producing the movie.

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

20. `spoken_languages`: The languages spoken in the movie.

21. `keywords`: Important terms or phrases associated with the movie, often used for categorization or search.

```
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 the last characteristic, `keyword`, from the data set due to its redundancy, irrelevance to the research question and presence of a significant number of null values. So, our data set now contains 20 variables.

Data processing

- Description of data processing you need to do to prepare for analysis, such as joining multiple data sets, handling missing data, etc.

For the following Imdb movies dataset, to prepare the dataset 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 quantiative 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 categorical predictors, such as adult and original language, need to be converted to factors to allow us to use them in our regression analysis.

- Visualizations, summary statistics, and narrative to describe the distribution of the **revenue** variable.

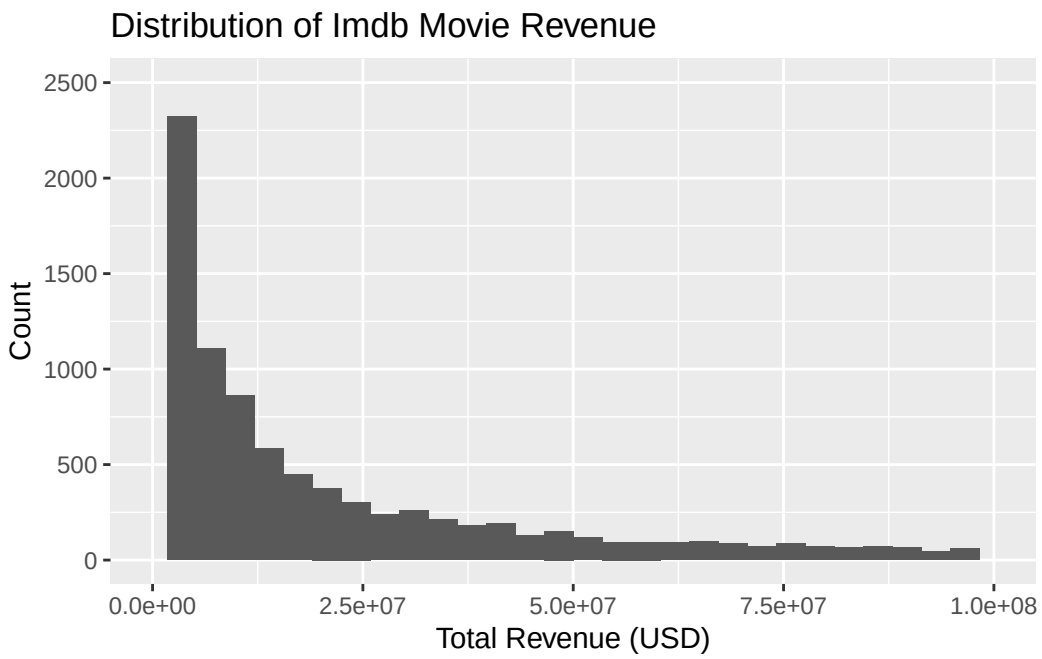
```
summary(imdb_data$revenue)
```

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

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

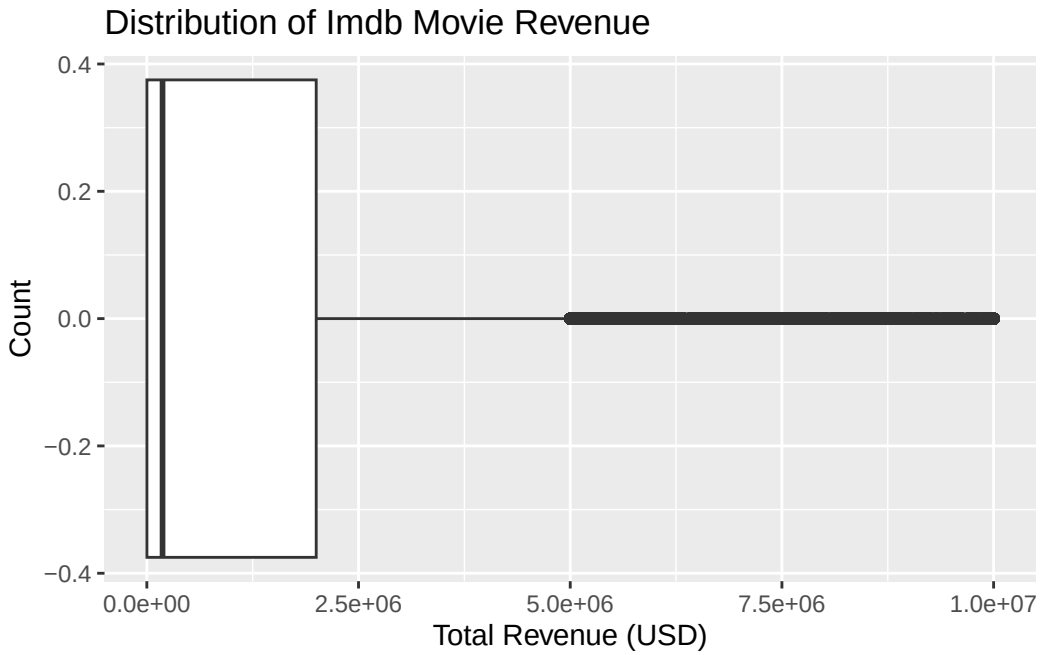
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
1.000e+00 2.400e+04 1.818e+06 3.857e+07 1.864e+07 5.000e+09
```

```r
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))
```




```r
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 dataset).

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 - Hellen

- A description of the potential predictor variables of interest
- Regression model technique (multiple linear regression for quantitative response variable or logistic regression for a categorical response variable)

Data Dictionary - Hellen

The data dictionary can be found here [Update the link and remove this note!]