**DATA**

* DESCRIBE THE DATA USED TO ANSWER THE QUESTIONS.
* EXPLAIN WHERE THE DATA ORIGINATED. FOR EXAMPLE, IT IS GOOD TO KNOW WHO COLLECTED THE DATA. JUST BECAUSE THE DATA CAME FROM KAGGLE, DOESN’T MEAN KAGGLE.COM COLLECTED THE DATA.
* IN-DEPTH DESCRIPTION OF THE SPECIFIC VARIABLES IN THE DATA REQUIRED TO ANSWER YOUR QUESTIONS.
* EXPLAIN WHAT EACH OBSERVATION REPRESENTS (I.E. PEOPLE, SCHOOLS, STATES, CITIES, PATIENTS FROM A SPECIFIC HOSPITAL).
* WHAT IS THIS A SAMPLE OF? HOW MANY OBSERVATIONS DO YOU HAVE?
* READER SHOULD CLEARLY UNDERSTAND THE SOURCE AND CONTENT OF THE DATA YOU PLAN ON UTILIZING TO ANSWER YOUR QUESTIONS THAT YOU PROPOSED IN THE INTRODUCTION.
* AT LEAST ONE, DESCRIPTIVE TABLE AND AT LEAST ONE FIGURE SHOULD BE USED HERE TO HELP THE READER UNDERSTAND WHAT THE DATA LOOKS LIKE WITHOUT SEEING THE ENTIRE DATASET.
  + IN ALL FIGURES AND TABLES, ONLY THE VARIABLES OF INTEREST SHOULD BE USED.

The dataset from Kaggle originated from TMDb (The Movie Database), where metadata from IMDB is scrapped by an international community since 2008 (<https://www.themoviedb.org/about>). The data was spread across two tables: movies and credits, however, we decided to only use the movies table. This table contained 4803 observations of movies, along with 20 columns regarding information of a movie’s title, keywords, revenue, and etc.

However, in this project, we will be focusing on only the variables listed below as well provided a description of each:

* Genre – the genres/categories of the movie (Action, Comedy, Fantasy, etc.)
* Original Language – the original language in which the movie was made
* Production Companies – the names of production companies producing the movie
* Tagline – the tagline of the movie
* Budget – the amount spent making the movie
* Popularity – a metric based on factors, such as number of votes for the day, number of views for the day, number of users who marked it as a "favorite" for the day, number of users who added it to their "watchlist" for the day, release date, number of total votes, previous days score to determine a movie’s popularity
* Release Date – the date the movie was released
* Revenue – the worldwide amount earned from the movie
* Runtime – the duration of the movie in minutes

The Production Companies and Genres variables were in JSON format so had to be converted into separate tables for each movie and its id. Since there were numerous production companies and genres, we selected the top 5-7 from each and created binary variables. We combined this data back to the original dataset in order to create one tidy dataset with 1434 complete observations and 22 variables.