# Project Proposal

Kable Guys: Jerry Hou, Arjun Prabhakar, Nathan Huang, Ryan Mitchell

October 28, 2021

## Section 1: Introduction

We are investigating the representation of women in films based on a test called the Bechdel test. For a film to pass the Bechdel test, it must have at least two female actresses who talk to each other about a topic besides a male character ('Call it the Bechdel-Wallace Test', The Atlantic). This simple test, coined by a cartoonist in a 1985 comic, helps us gauge the level of gender equality in a performative work (The Atlantic). Unsurprisingly, many of the blockbuster films fail the Bechdel test because Hollywood is infamous for lacking diversity in casts.

Since our dataset includes rating and grossing data for films, we want to take a closer look at the association between passing the Bechdel test and these performance metrics. A New York Times analysis of scores from Metacritic indicated that films directed by women of color were the best received by critics ('More Women Than Ever Are Directing Major Films, Study Says', New York Times). Additionally, several of the best-grossing movies from 2020 had predominantly female casts, so an analysis of a greater number of films can give us a better picture of the relationship between passage of the Bechdel test and a film's success.

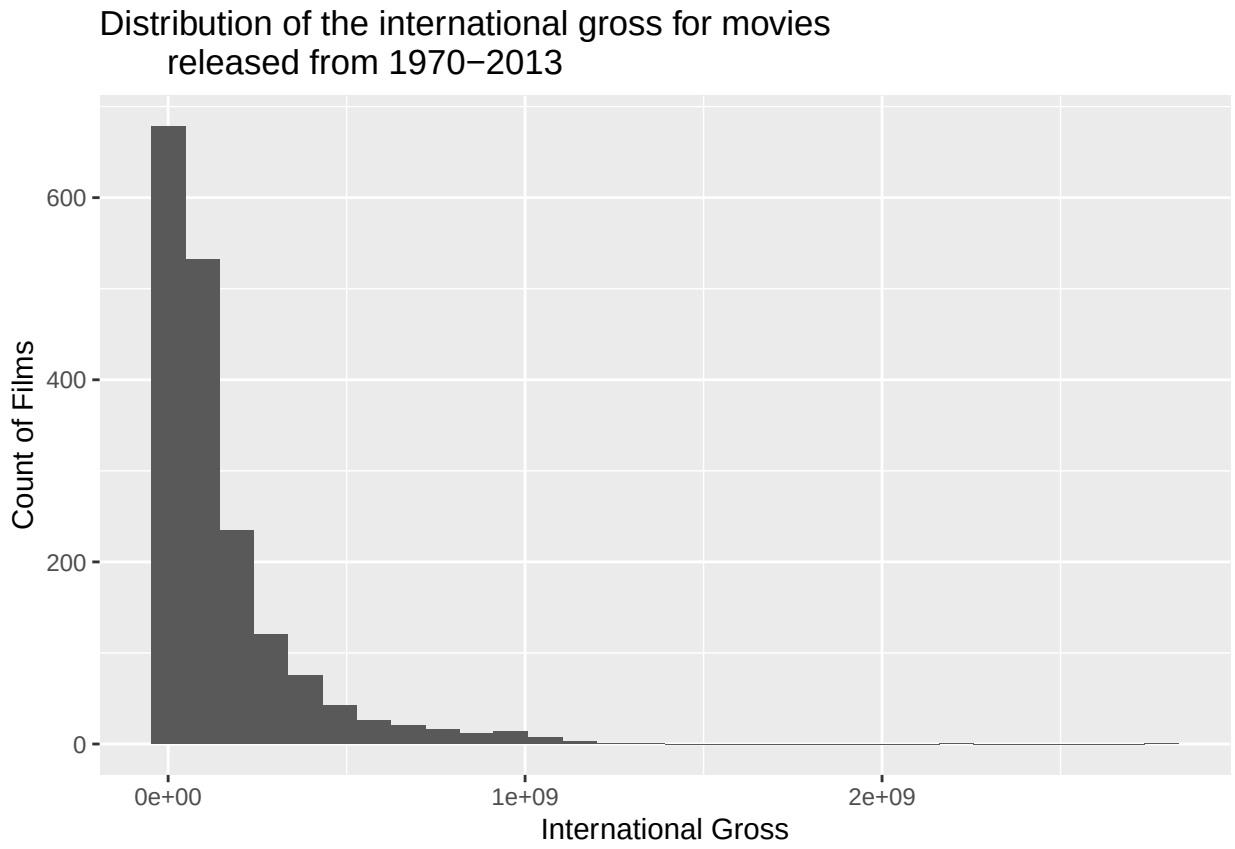
The general research question that we wish to explore is: Are films that pass the Bechdel Test better received by audiences and critics in terms of grossing and rating (measured by IMDB rating and international grossing variables)? Our hypothesis is that films that pass the Bechdel test tend to have a statistically significantly higher IMDB rating and international grossing amount than films that do not pass the test.

## Section 2: Data Description

The dataset used for this analysis is a part of the Tidy Tuesday repository, a collection of datasets released to the general public in an effort to produce varied data visualizations. The dataset comes from a FiveThirtyEight article titled "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women," which relied on data from BechdelTest.com and The-Numbers.com. FiveThirtyEight used the former to determine if films passed the Bechdel Test and the latter to obtain financial information on the films. The `movies.csv` dataset has 1794 observations, each representing a film released from the years 1970 to 2013, and 34 variables. Key variables in the dataset are genre, IMDB rating, domestic gross, international gross, budget, awards, and whether or not the film passes the Bechdel Test. To answer our research question, we will mainly take a look at the `binary`, `intgross`, and `imdb_rating` variables.

## Section 3: Analysis Approach

Our response variable will be the international gross, which is a quantitative variable representing how much the movie the observation represents made worldwide.




| mean_int_gross | median_int_gross | sd_int_gross | IQR |
|---|---|---|---|
| 150385700 | 76482461 | 210335268 | 163721382 |

The distribution of international gross is heavily skewed right. We can estimate the center of the data with the median, which is $76,482,461, and we can estimate the spread with the IQR which ranges $163,721,382. The international grosses range from $828 to $2,783,918,982.

We consider the following variables to be predictors:

- `binary` - Whether a film passes the Bechdel Test
- `year` - Year of the movie's release
- `budget` - Budget as of release year
- `type` - Type of Film
- `genre` - Genre of the movie

Since our response variable, the gross international revenue, is continuous, we will use a multiple linear regression model. Because `intgross`, the international gross revenue, is not normal and has a strong right skew, we look to use a multiple linear regression model with the log(`intgross`) values using the aforementioned predictor variables.