# MA 592 Project Proposal

Zhiwei Liang, Yijia Wei, Yuankai Zhang

# Introduction

The market size measured by revenue of the Movie Production industry in the US is 17.4 Billion in 2021. It is important to predict movie rating and revenue for planning both production and distribution strategy. Some studies have shown that many factors are related to the success of movies, such as budget, rating, release date, star-power, and so forth. However, little is known about the causal effect of the budget on the movie rating. In this project, we propose to examine the causal effect of budget on the average rating of movies collected from The Movie Database.

# Causal Question

**What is the average rating of the movie if all the movies have a high-level budget?**

* treatment: budget level;
* outcome: average rating;
* confounders: production countries, genre, original language, released date period

# Data

**Data Description**

Our dataset contains metadata for all 45,000 movies in the Full MovieLens Dataset (before 2017). It’s convenient to reselect the features and update the content. For each movie, there are 16 variables included as the following:

* **budget**: The budget of the movie in dollars.
* **genres**: A stringified list of dictionaries that list out all the genres associated with the movie.
* **belongs\_to\_collection:** if the movies belong to a series or a collections
* **id**: The ID of the move.
* **original\_language**: The language in which the movie was originally shot in.
* **original\_title**: The original title of the movie.
* **overview**: A brief blurb of the movie.
* **popularity**: The Popularity Score assigned by TMDB.
* **production\_companies**: A stringified list of production companies involved with the making of the movie.
* **production\_countries**: A stringified list of countries where the movie was shot/produced in.
* **release\_date**: Theatrical Release Date of the movie.
* **revenue**: The total revenue of the movie in dollars.
* **runtime**: The runtime of the movie in minutes.
* **spoken\_languages**: A stringified list of spoken languages in the film.
* **title**: The Official Title of the movie.
* **vote\_average**: The average rating of the movie.
* **vote\_count**: The number of votes by users, as counted by TMDB.

**Data Collection**

We gain this dataset from [Kaggle](https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv), and it is originally collected from Full MoviLens. We also process the data cleaning, since there is missing data and some variables are in JSON format.

# Model

1. Linear Regression
2. Logistic Regression
3. Conditioning on baseline covariate
4. Inverse Probability of Treatment Weighting
5. Doubly Robust
6. Propensity Score
7. …...

Reference

1. Ibisworld.com. 2021. IBISWorld - Industry Market Research, Reports, and Statistics.
2. Cho, S., Kim, H., Kim, B. and Kim, H., 2014. Predicting Movie Revenue by Online Review Mining : Using the Opening Week Online Review. Information Systems Review, 16(3), pp.113-134.
3. Nikhil Apte, Mats Forssell, Anahita Sidhwa :“Predicting Movie Revenue”, CS229, Stanford University ,December 16,2011