Springboard Capstone
John Sizemore
sizemore.fox@gmail.com

Background

Every industry contains a certain set of unwritten rules or stories that identify standard behaviors and set out what it takes to make money. In the film industry, one of the most common and pervasive of these rules is that the season in which a film is released has an outsized effect on the chances of its success. Studios have created stories (as in mental narratives that state a perspective on how the world works) about consumer behaviors and the relationship between those behaviors and the probability of success for any particular film. Stories about when children are in or out of school or whether or not consumers (at large) are traveling for summer or winter holidays have a very real impact on studio executives as they plan the release of a film or slate of films.

These stories, however, aren't necessarily evaluated for their validity; they are the result of years and years of experience releasing movies. This is not to say that this experience isn't valuable; it is beyond a doubt valuable. Further, these stories are reinforced by reality; since a studio will only try to release a film they expect to be a hit during a window in which they think it will become one, hits generally occur (or are perceived to occur) during those windows. This is why we see many big budget, super-hero movies released back to back in early summer and none released during the month of September.

Recently, there have been some counter-examples to this trend. The most prominent of these counter-examples is likely Deadpool, which was released Feb 12, 2016. Typically, a blockbuster comic book movie would be slated for the summer or the holidays and would avoid releasing on or around something like Valentine's Day (or in the lead up to Spring Break). Despite the 'headwinds', Deadpool greatly outperformed expectations, bringing in more than $750M in worldwide box office.

What I'm Trying to Solve

Ultimately, the phenomenon observed with regard to releasing films becomes a 'chicken or the egg' situation: are the "prime" months for releasing a film (May, June, July, Nov, and Dec) actually more conducive to success because consumers want to see movies more during those times than at other points of the year, or is it a result of studios only releasing movies they believe to be potential hits during those months? Is the phenomenon a result of supply (when studios release films) or demand (when consumers are most willing to go the movies)? Are these "prime" windows even more indicative of commercial success?

Strictly speaking, I am not trying to create model to predict movies, nor I am trying to understand what makes a movie a smash-hit. As works of art, films have a great many nuances that are difficult to capture with publicly available data. I am only concerned with investigating a sacred cow/dogmatic belief (namely, that there are times of year consumers do not want to see movies) to understand whether or not there is a data-driven reason to continue to believe in it, or if in fact, there is an advantage to be gained by being the prime mover in discarding this belief.

The Datasets

For this analysis, I will use two primary datasets. First, I use historical box office information for every movie released between 1970 and 2017. For ease of communication, I'll refer to this data set as the movies dataset. Second, I have located a dataset from the Bureau of Labor Statistics (BLS) about an annual and ongoing survey they perform investigating how Americans spend their time (i.e. the American Time Use Survey). For simplicity, I will refer to this dataset as the ATUS.

The Movies Dataset

The movies dataset contains historical summary information about films and is sourced from a data-distributor called Opus Data (the company and data behind the movie website the-numbers.com). Opus provides a web interface for querying the database, so much of the filtering I had to do was done outside of R via the Opus portal. I only wanted to include movies for which I had budget information and that made at least some significant amount of money in theaters (i.e. total box office < $1,000,000). These filters created a dataset of 5,003 films. A quick summary of the pertinent fields is below:

Field	**Description**
Title	Film title
Is_Sequel	Whether or not the film is a sequel (or more generally a part of a franchise)
Run time	The length of a film
Source	Where the idea/source material for the film originated (e.g. if the film is an adaptation of book)
Creative type	What kind of story is it (e.g. Historical fiction, science fiction, Dramatization, Super hero, etc.)
Production method	What kind of production is it, live-action, animation, etc.
Genre	The genre of the film (e.g. Horror, Sci-fi, Comedy, Action/Adventure)
Production budget	The amount of movie the studio reported spending to produce

	the film (not inclusive of marketing or other distribution costs)
Domestic box office	The amount of money the film made in theaters in the United States and Canada
International box office	The amount of money the film made in theaters in any country that is not the US or Canada
Total box office	The sum of Domestic and International box office figures
Inflation adjusted domestic box office	Opus Data provides an inflation adjusted figure for the domestic box office
MPAA Rating	The film's rating from the Movie Picture Association of America (e.g. R, PG-13, PG, etc.)
Theatrical Release Date	The date on which the film was released in theaters (typically for domestic release)
Opening Theaters	The number of theaters in which the film opened domestically

The ATUS dataset

This data was downloaded directly from the BLS website at www.BLS.gov/data. The ATUS data used in the analysis is pulled from two separate tables. One of the tables is the respondent level data (call Surv_resp in my code), which is essentially a daily diary completed by participants in the survey. During the months of participation, each participant filled out a diary detailing her activities. The pertinent fields from this data set are:

Field	**Description**
TUCASEID	A unique identifier for each response in the survey
TUACTDUR	The duration reported for any particular activity in the daily diary
TRCODEP	The activity code for the reported activity (codes created by BLS for the survey). For our purposes, we are concerned only with TRCODEPs 120303 (time spent watching TV or movies at home) and 120403 (time spent watching movies in theaters)

The other table from this survey that is used is the current population survey component of the ATUS. This table includes information about the participants and their households. The pertinent fields for this analysis are:

Field	**Description**
TUCASEID	A unique identifier for each response in the survey
HRMONTH	The month in which the participants responses occur
HRYEAR4	The year in which the participants responses occur

Data Wrangling

As both datasets came from curated online repositories, much of the data wrangling was already handled. For the most part, I just needed to create additional variables that would enable my analysis.

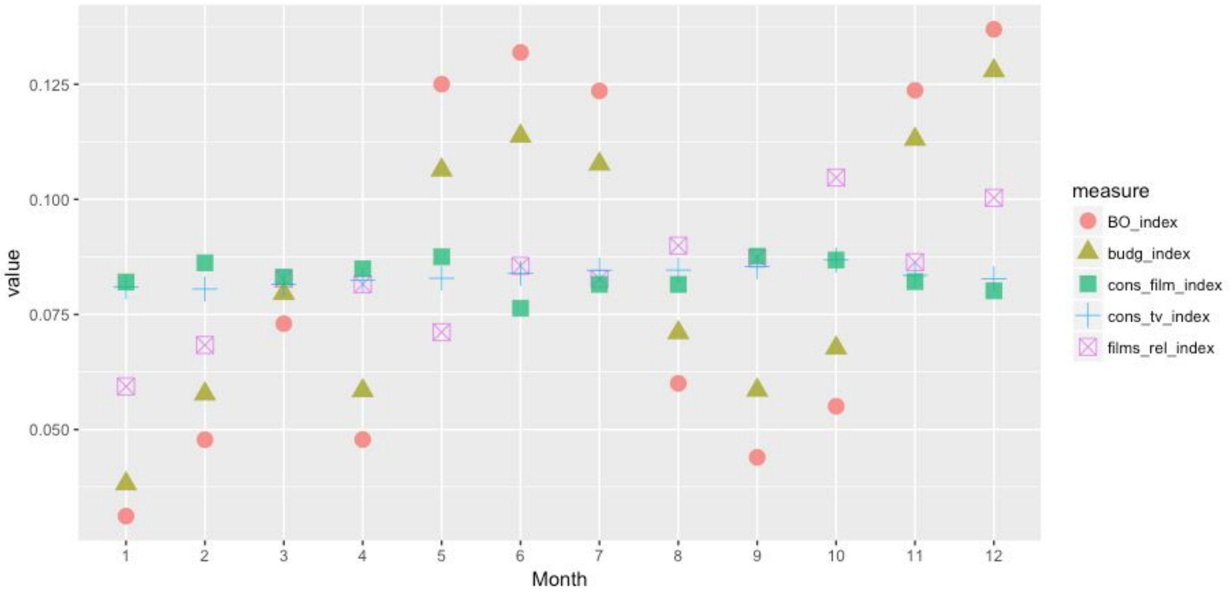
First, I needed to create a metric that most closely captured the outcome I am looking to investigate. Total box office is a fine metric for the success of films, after all, movies are supposed to make money. But, box office doesn't account for cost or consider that rate of return could be more important (i.e. could you make more money making ten $10M movies than one $100M movie?). So, I calculated a performance ratio for each film, which is total box office divided by production budget. Since I am interested in evaluating the effect of seasonality on whether or not a film is successful, I had to come up with a threshold of performance ratio that determines success. After some trial and error, I chose a threshold of 4; though unscientific this threshold does create a universe in which ~23% of films are considered commercially successful, which through experience I can say is close enough to right for our purposes. Using this threshold I created a binary variable called Is_comm_success and assigned 1 to films exceeding the threshold and 0 to those that did not.

In order to most easily compare films over time, I also needed to adjust box office and production budget numbers for inflation. To do this, I used the Federal Reserve Bank of St. Louis website to pull inflation figures (Consumer Price Index) for 1970 to 2017. I indexed the data for the last entry in Dec 2017 (Infl_index in the dataset). I then calculated an inflation adjusted box office and inflation adjusted production budget for each film (Infl_Dom_BO_FRED and Infl_adj_prod_budg, respectively). I ignored the inflation adjusted box office figures I received from Opus Data to ensure consistency in the calculation/indexing.

Analysis - Seasonality of Box Office and Consumer Behavior

For the purposes of all of these analyses, I use production budget as proxy for gauging how a studio perceives the potential profitability of film. As with any business, a studio wouldn't normally go about spending more to make a product than they think it will make back when they go to sell it. This is admittedly a flawed metric, as there are certainly times when a studio does make a movie they do not believe will be commercially viable. The reasons for this are (typically) either that the studio believes the film can win major awards (like an Oscar) or the studio wants to develop or maintain a relationship that is important (for example, a studio might make a less expensive passion project for an actor or director in order to convince that actor or director to make their next big film with that studio). All that said, I think using budget as proxy for studio perceptions of potential works just fine, given both that it is pretty much the only public information available and also that studios must care about the budget as it relates to potential performance in the vast majority of cases.

To begin the analysis, I decided to first look at the film market in aggregate. Consolidating all the information on a month by month basis, I looked at the breakdown of consumer behaviors, box office returns, production budgets, and overall supply of films. The resulting chart looks like this:



I indexed these summarized variables by calculating the percentage of the entire population represented by each month. For example, we can see that approximately 2% of all box office returns, 3% of all production budgets, 7% of all films released, and roughly 8.5% of the time consumers reported watching TV and movies occured in January.

That we observe a tight correlation between box office returns and budget is to be expected; movies that cost more money to make should make more money in the theaters. The interesting thing is to note how this chart displays the seasonality of box office returns, but no corresponding seasonality is observed in either of the metrics about consumer behaviors. From an aggregate perspective, over the course of the 13 years of the ATUS study, consumers report watching TV and going to the movies at pretty much the same rate throughout the year. There isn't an observable increase in consumer consumption in the high box office months of May, June, July, November, and December.