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. As an example, of the 102 G-rated films in the dataset I used for the analysis below, zero were released in the month of September. Obviously, there is some pervasive story about the 'back to school' month that prevents studios from releasing family films in that month.

These stories, however, aren't necessarily evaluated for their validity; they are the result of decades of releasing movies in this pattern. 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, super-hero, based on a 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, 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 more indicative of commercial success? If so, is that the result of consumer behaviors or preferences?

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 dogmatic belief (namely, that there are times of year consumers do not want to or will not go to see movies) to understand whether or not there is a data-driven reason to continue to believe 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 response occurs
HRYEAR4	The year in which the response occurs

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 (perf_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. If a film makes four times its production budget, it likely recouped all marketing and distribution costs and provided a relatively decent return to the studio and its financiers. 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

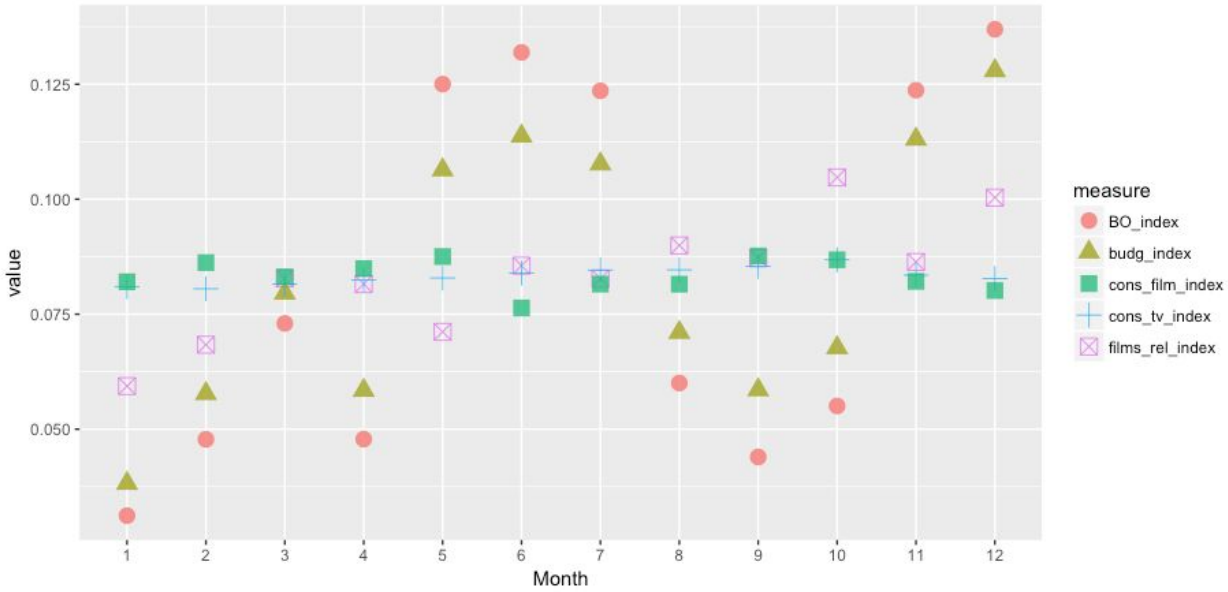
adjusted box office and inflation adjusted production budget for each film (Infl_Dom_BO_FRED and Infl_adj_prod_budg, respectively). I ignored the inflation adjusted box office figures I received from Opus Data to ensure consistency in the calculation/indexing.

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.

Aggregate Market Analysis

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:

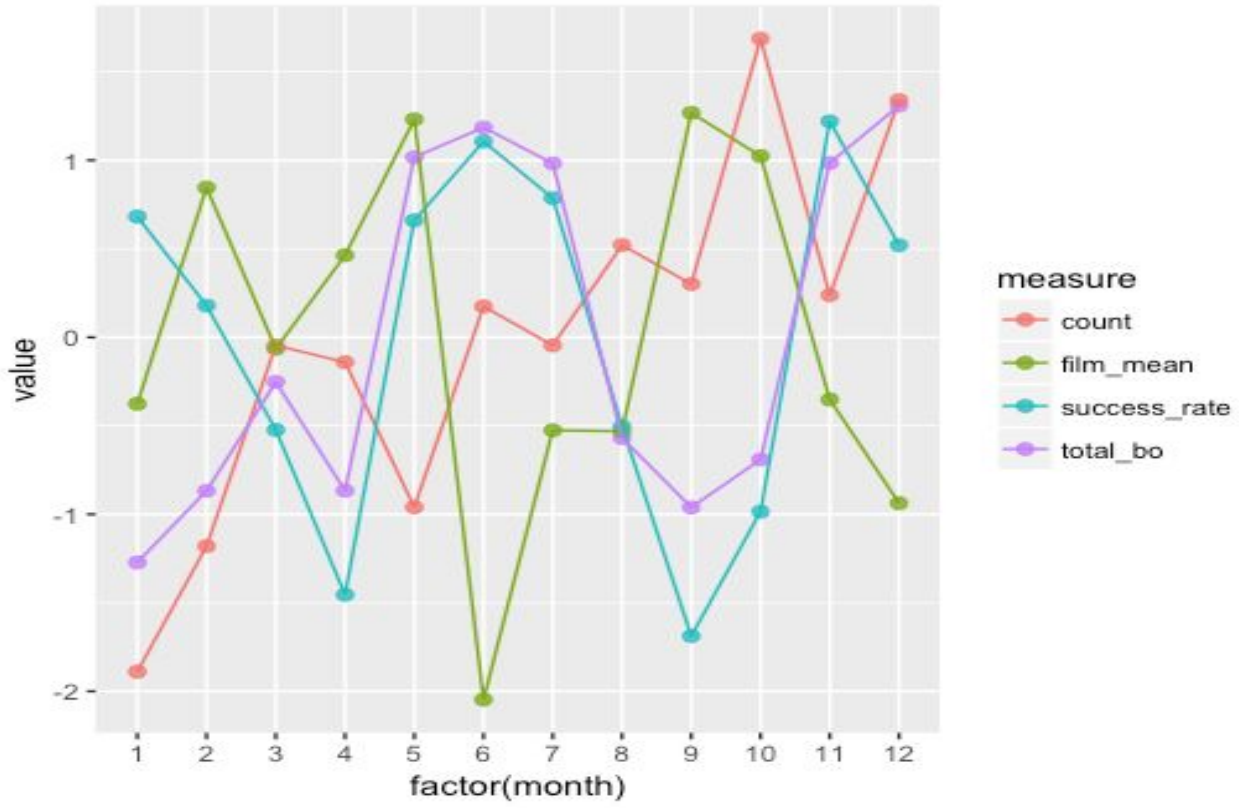


[For this chart, BO_index is box office returns, budg_index is production budget, cons_film_index is the time consumers spent watching movies at theaters, cons_tv_index is the time consumers spent watching TV at home, and films_rel_index is the number of films released by month.]

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.

The next step in my analysis was to examine success rate, as opposed to box office. As I noted above, I am more interested in the effect of seasonality on the success of a film, not just on total box office. To that end, I included success rate in the chart from above (and removed both budg_index and TV_index to make it easier to read). I also changed the index calculation by using scale() instead of percentage of population. The resulting chart is below:



[For the above, count is a count of films released, film_mean is the time consumers spent going to the movies, success_rate is the percentage of films in that month that are commercially successful, and total_bo is the sum of all box office figures.]

Looking primarily at the film_mean (the amount of time consumers reported spending watching films in theater), there is a strange relationship with both success rate and total box office. It appears as though the months in which people report spending less time in theaters are actually the months in which the success rate is higher than average and vice versa. Another takeaway from the above chart is that consumer behavior actually runs counter to the established narrative. The entertainment industry believes people spend more time at theaters in the summer and during the holidays, but the chart shows the opposite is true. While May is a strong performer, June, July, Nov, and Dec are 4 of the 5 worst performing months when it comes to the ATUS data.

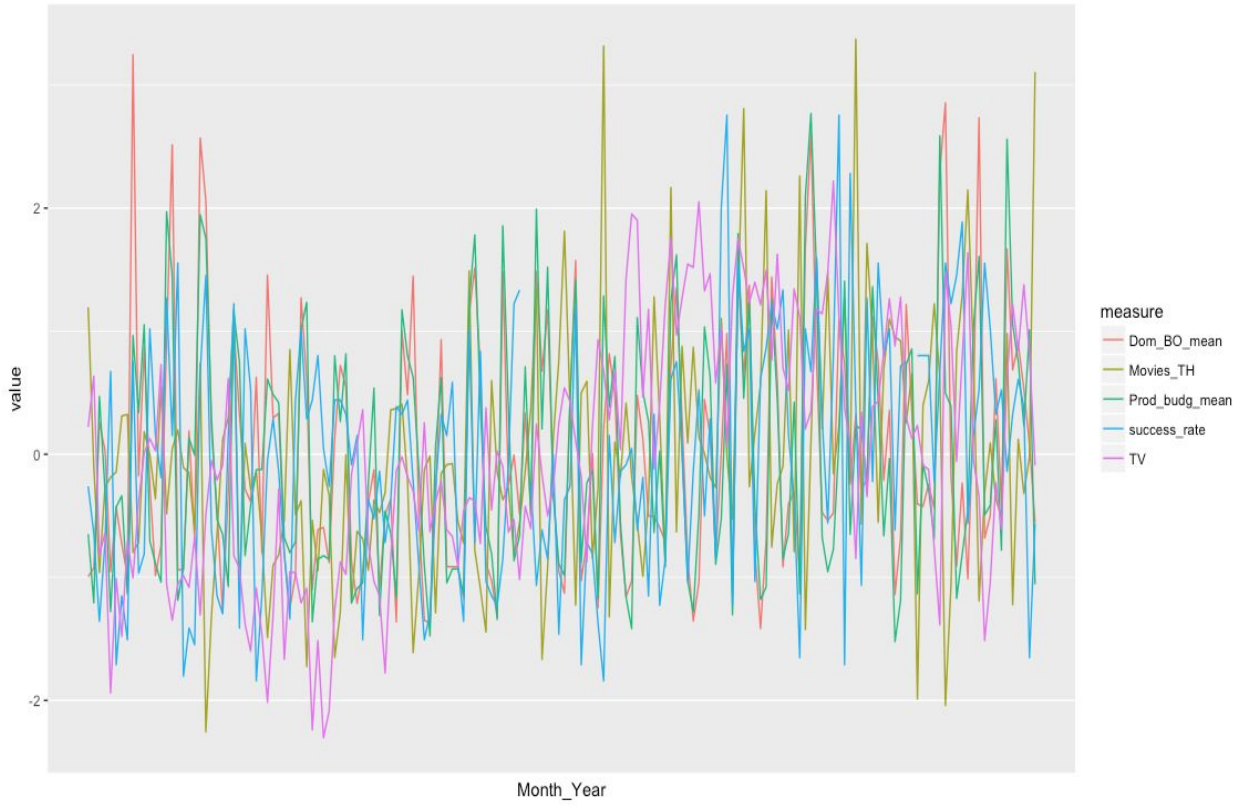
These observations are counter-intuitive to say the least, but there are potential explanations. First, it is possible that the survey doesn't capture some subset of the population that sees more movies in certain months as opposed to others (e.g. children). Second, it is useful to consider that the absolute difference here is actually quite small, though the indexed difference is large. The standard deviation of the film_mean in this sample is only ~5.5 minutes per month. So, the amount of actual difference between April and May consumption numbers is only about 15 minutes (or a fraction of a film), and it's at least possible that the duration of films in April is longer than in May.

Additionally, it appears as though film supply has an effect on success rate. Take January as an example. There are far fewer films released in January than any other month, however, January's success rate is much higher than average. Consumers report approximately average amount of time at the theater, but since supply is lower, success rate is higher. October is another interesting point for considering supply. October has the greatest number of films released, but the third worst success rate, despite consumers reporting higher than average time in the theater. People spending more time in theaters, but diluting that time across a larger supply of movies leads to a lower rate of success. Months in which supply is either constrained (like January) or the market is saturated (like October) under or over perform.

Year-month Market Analysis

Next, I decided to look at the data on a more granular level. Instead of aggregating by month over all years, I aggregated the movies dataset and the ATUS data set at the year and month level. For the ATUS survey, I only had data for Sept. 2002 through October of 2016, so the following analysis is limited to that scope of time.

Similar to above, I aggregated box office, production budget, consumer time spent watching TV, consumer time spent seeing films in theaters, and success rate for each month in the sample (170 months). Once again, I used scale() to normalize the data, and created the following chart:



Not much to be gleaned from looking at the above, so instead I looked at a correlation matrix of the variables:

	Movies_TH	TV	Prod_budg_mean	Dom_BO_mean	success_rate
Movies_TH	1.000000e+00	0.26568657	-0.09057145	-0.1535950	7.546252e-05
TV	2.656866e-01	1.00000000	0.01318528	-0.0131786	8.166689e-02
Prod_budg_mean	-9.057145e-02	0.01318528	1.00000000	0.8780118	1.798313e-01
Dom_BO_mean	-1.535950e-01	-0.01317860	0.87801179	1.0000000	3.519357e-01
success_rate	7.546252e-05	0.08166689	0.17983131	0.3519357	1.000000e+00

Once again, we see that there isn't a meaningful relationship between any of the variables and success rate. The strongest correlation in the set is between production budget and box office, a phenomenon I have already covered above. The next strongest correlation is between domestic box office and success rate, which makes pretty obvious sense. The more money a film makes, the more likely it is to become a commercial success.

Film by Film Analysis

The final step in my analysis was to run a logistic regression on all the movies that fall in the window during which I have consumer survey information. To do that, I filtered the movies data set down to just films released between Sept. 2002 and Oct. 2016 (a sample of some 3,179 films). Then, I added the results of the consumer survey for the month of release to the entry for each individual film. I also calculated and added a variable for how many films were released in the same month as the film in question. I removed all variables unrelated to this analysis and ended up with a dataset that was the title of the films, whether or not it was 'commercially successful', the month of release, how many films were released in the same month, and the production budget (adjusted for inflation).

To prepare for the logistic regression, I created dummy variables for the month of release (a categorical variables), and I normalized the numerical variables using scale(). I ran the logistic regression and ended up with these results:

```
Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.371874   0.151037  -9.083   <2e-16 ***
Rel_month1         0.307869   0.231645   1.329   0.1838
Rel_month2         0.176264   0.225908   0.780   0.4352
Rel_month3        -0.190512   0.224565  -0.848   0.3962
Rel_month4        -0.390276   0.229007  -1.704   0.0883 .
Rel_month5         0.196819   0.230731   0.853   0.3936
Rel_month6         0.252360   0.213157   1.184   0.2364
Rel_month7         0.199608   0.211333   0.945   0.3449
Rel_month8        -0.018026   0.212738  -0.085   0.9325
Rel_month9        -0.212913   0.218138  -0.976   0.3290
Rel_month10       -0.039073   0.213703  -0.183   0.8549
Rel_month11        0.208227   0.211920   0.983   0.3258
Rel_month12              NA         NA      NA       NA
Infl_prod_budget1 -0.009058   0.045095  -0.201   0.8408
Movies_TH1         0.007143   0.048003   0.149   0.8817
YM_count1         -0.033063   0.056144  -0.589   0.5559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3220.2  on 3148  degrees of freedom
Residual deviance: 3196.3  on 3134  degrees of freedom
AIC: 3226.3
```

As is obvious, there isn’t a clear relationship between any of these variables and commercial success (which is oddly what I was hoping for). Month release, production budget, supply (or movies released in the same month), and the amount of time consumers report spending going to the movies are not statistically significant when it comes to predicting individual film success.

However, I ran a subsequent logistic regression in which I collapsed all the months down into a simple binary variable, Is_rel_prime, which captures whether or not a specific film is released in May, June, July, November, or December. The summary of this model is:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.4248702  0.0594225 -23.979   <2e-16 ***
Is_rel_prime      0.2062711  0.0945095   2.183   0.0291 *
Infl_prod_budget -0.0146759  0.0448819  -0.327   0.7437
Movies_TH         0.0006479  0.0444914   0.015   0.9884
YM_count         -0.0894518  0.0464080  -1.928   0.0539 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3220.2  on 3148  degrees of freedom
Residual deviance: 3209.5  on 3144  degrees of freedom
AIC: 3219.5
```

This model suggests that there is in fact a positive relationship between prime months and commercial success, as well as a negative relationship between the number of films in the market and commercial success. Both of these findings are not surprising, and in my opinion, they are not rebuttals of the point that consumer demand doesn't necessarily change with the seasons. They support the conclusion that the things the studios control have more influence than the things they do not.

Conclusion

From all of the above, it would be equally unfair to conclude that A) consumer behaviors and preferences play no role in the seasonality witnessed in box office returns (in the end, box office returns are not separate from but are in fact a consumer behavior), or B) that that consumer appetite for entertainment alone drives seasonality. In aggregate, consumers do not report spending substantially more or less time consuming entertainment (either at home or at the theater) through the course of the year. From the ATUS data, we can reasonably state that consumer demand for entertainment is (mostly) consistent; people want to be entertained regardless of what the calendar says.

It appears that just as strong a case can be made that studio decisions about when to release the films it thinks have the highest potential (as measured by budget) has a greater effect on the overall market (and to a lesser extent on success rate). Further, the data (and common sense) suggests that months in which there are fewer films, aka months in which studios constrain supply, have higher success rates vs months where there are more films, aka when the market is oversaturated.

All of this suggests that a studio should be more concerned with appropriately gauging opportunity (thereby controlling budgets) and with what the competitive landscape is than with which month a film should be released. Considering that people are willing to spend about the same amount of time consuming entertainment, finding times when there isn't much competition for that same amount of time could lead to an increase in performance or probability of achieving commercial success. Obviously, this isn't the only factor at play, but the data suggests it should be in the conversation.

What's Next

This analysis falls short in several ways. First and foremost, there are peculiarities around the ATUS survey that deserve a much closer look. For example, that consumers report spending the least amount of time watching movies in theaters in the month of June, yet June boasts both high gross box office and a high success rate doesn't really make sense. There is a chance that there is some bias in the sample (for example, children are excluded from the survey, but parents buy movie tickets for children), or there is a chance some labels/codes have been misinterpreted by me (the documentation for this study is literally hundreds of pages long). Given the time, one of the first next steps to take would be to either go deep on the ATUS survey, or to find another data set that would help identify/understand the patterns of consumer behavior across the year. Certainly, the ATUS data would have some interesting insight if much of the demographic information is included (and understood).

Next, the movies data and the ATUS data is limited in its granularity. Given daily surveys and/or daily/weekly box office numbers, one could have a much better understand the relationship between consumer preferences, the calendar, the supply of movies, and box office performance. Using more granular data would allow us to understand the effects on performance of new entrants into the market, as well as how the new effects vary by different groups of consumers (when the above-mentioned demographic information is included).

Further, this analysis presumes that all films are created equal, when that is not necessarily the case. Films are made for different motives and can have wildly different characteristics (budgets, ratings, genres, and so on). It is entirely possible that seasonality would be more pronounced in either the movie or the ATUS dataset if smaller sub-groups were considered (or particularly if target audiences could be identified). For example, if we could identify the survey participants with children who would be the target audience for animation films, we might observe that they spend a greater amount of time seeing films in summer months than at other times, thereby justifying studio-imposed seasonality of animated films.

Finally, it's really a negative case for which I'm arguing -- saying that the "prime" season for movies isn't purely a function of consumer preference isn't as strong as saying "prime" season doesn't exist. "Prime" season does clearly exist (insomuch as films released in May, June, July, November, or December do have higher than average box office numbers and success rates), but proving that a film released in that window would have performed differently had it been

released at some other time is impossible. I think there are circumstances in which tests could be created (e.g. staggering international releases to see the effects of season), but these tests would be difficult to pull off and would likely not be performed on big-budget films, so the results may not be conclusive or helpful.