Springboard Capstone - Data Wrangling Exercise

Introduction:

It is common practice in the film industry to avoid releasing films during certain windows in the year (the months of August and September, March, etc.). These rules of thumb are based on stories about consumers' preferences, not on hard data. Recently, some films have 'broken' the rules and succeeded.

For example, Deadpool was released in February (typically avoided because Valentine's Day and approaching Spring Break season) but was a huge success. The purpose of my capstone is to examine if historical box office patterns justify the seasonality studios impose on movies, namely, does box office performance in those months suffer because people don't want to see movies at that time of year or because studios don't release films. Is the problem the supply of films or the demand from consumers?

Data Wrangling:

The primary data set for this analysis is a set of historical summary information about films released since 1970. Fortunately, the source for this data (a company called Opus Data) has a pretty robust web portal, so a good deal of the filtering (e.g. leaving out films for which there isn't budget information) is handled outside of R. That said, there were a few minor changes I needed to make which are detailed in the movies_wrangle.R script in my github repository.

There are a couple additional data sets I plan to use. These data sets are from the Bureau of Labor Statistics American Time Use survey. This data is from an annual survey done by the BLS asking Americans how they spend their time. The purpose of including this data is to use it as a proxy for any seasonality in how people consume entertainment. Again, this data is already pretty clean, just needs to be filtered for the appropriate analysis.

Datasets are too large to be stored on github, but are all available from the BLS website.