Box Office Mojo EDA

Oleg Cherkasskikh

9/8/2018

Data: BoxOfficeMojo.com All-Time Domestic Data  
Source: <https://www.kaggle.com/eliasdabbas/boxofficemojo-alltime-domestic-data/version/2>

Libs = c(tidyverse, readxl, treemapify, RColorBrewer)

Introduction

There is only 1 data frame file (boxoffice.csv), containing 5 variables:

1. **Rank**: Lifetime rank of the movie
2. **Title:** Name of the movie
3. **Studio**
4. **Lifetime\_gross**: domestic US sales, not adjusted for inflation
5. **Year**: year of production

Data Preparation

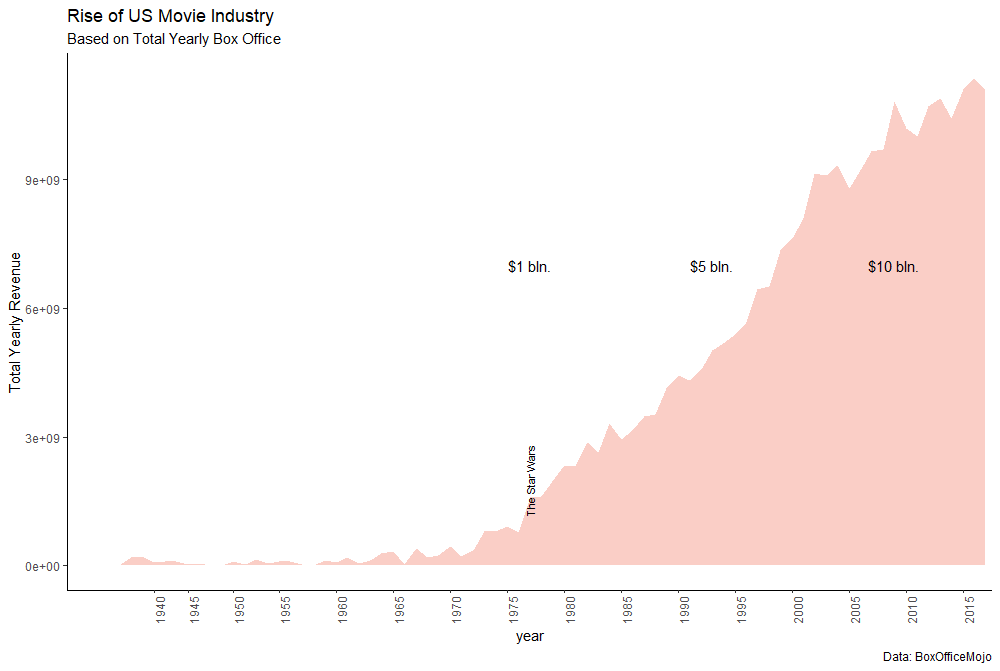
* Fields are mostly in ready form
* Computed “decade” column for future visualizations, by creating a binning if function and using dplyr::mutate().

Objective & Methodology

To perform EDA on a dataset with a limited number of variables and visualize the results that would provide insights into the trends and idiosyncrasies of the US movie industry since its inception.   
  
To avoid overcrowding, the R code for the graphs presented can be found in an attached pdf ‘BoxOfficeMojo Graphs’. The full source code can be found in the attached R file.

I. Rise of the American Movie Industry

Visualizing the total revenues by year for all studios allows us to gain perspective of the industry’s growth. As going to the movies hasn’t really become a part of America’s pastime in the 60’s, the directors began to really explore the creative boundaries by moving away from traditional Hollywood production code in the 70’s. The era of civil rights, the Watergate scandal, and the Vietnam War has left Americans longing for a shift in cinema offerings, moving toward sci-fi genre and “super-protagonist” oriented movies. The 1970’s served as a platform for multiple super successful movies franchises , such as Jaws, Alien, The Godfather, Superman, Westworld (the original) and the biggest one to date – Star Wars. It is the decade that established movie watching as America’s #1 leisure activity and allowed Hollywood to cross the mark into becoming a billion-dollar industry (as seen on the graph). As global demand scaled exponentially, and studio growth transformed them into the titans of entertainment, the industry crossed the 5 billion mark around 1993 (*Jurassic Park, Schindler’s List*), and 10 billion mark in 2009 following the box office success of James Cameron’s *Avatar* (the world’s highest grossing film of all time, and one of the first to take full advantage of the IMAX technology outside of the US).

II. Most Profitable Movies each year and each decade

Running a simple dplyr query allows us to look at the most popular movies each year and decade, which is beneficial if we want the blockbusters, but visualizing the trend of box office revenues doesn’t provide us with any meaningful insights, since it looks very similar to general industry growth. Therefore, only the snippets of data frames are presented below:

title year

1 Black Panther 2018

2 Star Wars: The Last Jedi 2017

3 Rogue One: A Star Wars Story 2016

4 Star Wars: The Force Awakens 2015

5 American Sniper 2014

# Most profitable movie for each **decade.**

title decade

*<chr>* *<fct>*

1 Star Wars: The Force Awakens 2010s

2 Avatar 2000s

3 Titanic 1990s

4 E.T.: The Extra-Terrestrial 1980s

5 Star Wars 1970s

6 The Sound of Music 1960s

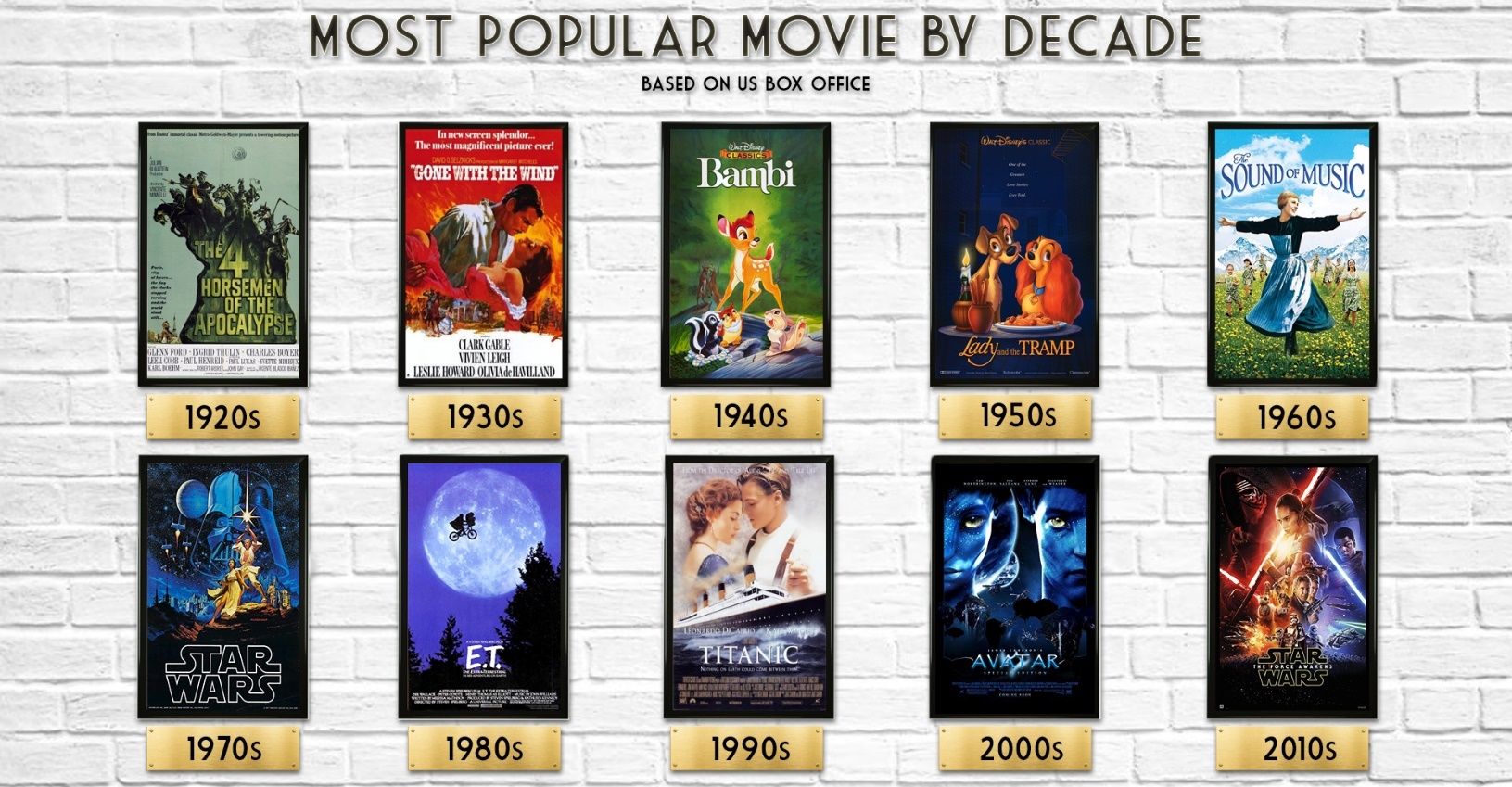
7 Lady and the Tramp 1950s

8 Bambi 1940s

9 Gone with the Wind 1930s

10 The Four Horsemen of the Apocalypse 1920s

I thought it would be most appropriate to do a little photoshopping to visualize this data as a poster wall. Not R related, but appropriate for visualization purposes (maybe…).



III. Do Movies with Shorter Names Tend to Be More Successful than The Longer Ones

It’s the question for the ages that is finally being answered - Do movies with shorter titles attract wider audience? Is there a psychological component that nudges the movie-goers toward movies with shorter names? While we don’t have proper data to answer the 2nd question, we can analyze the revenues, and come up with an educated explanation for the outcome.

First, I found out that the average number of characters in all the titles is ~14.7 units, and the median falls on 14. Hence, I wrote this query to separate movies into 2 categories: the ones with shorter names(<=14 characters), and the ones with longer names( >15).

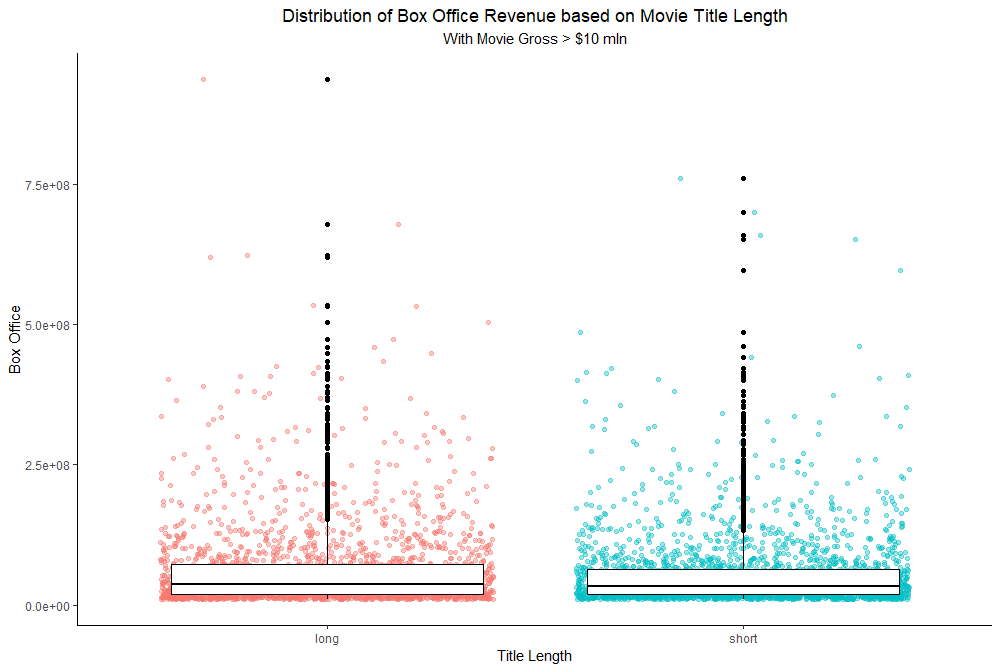
boxoffice %>%

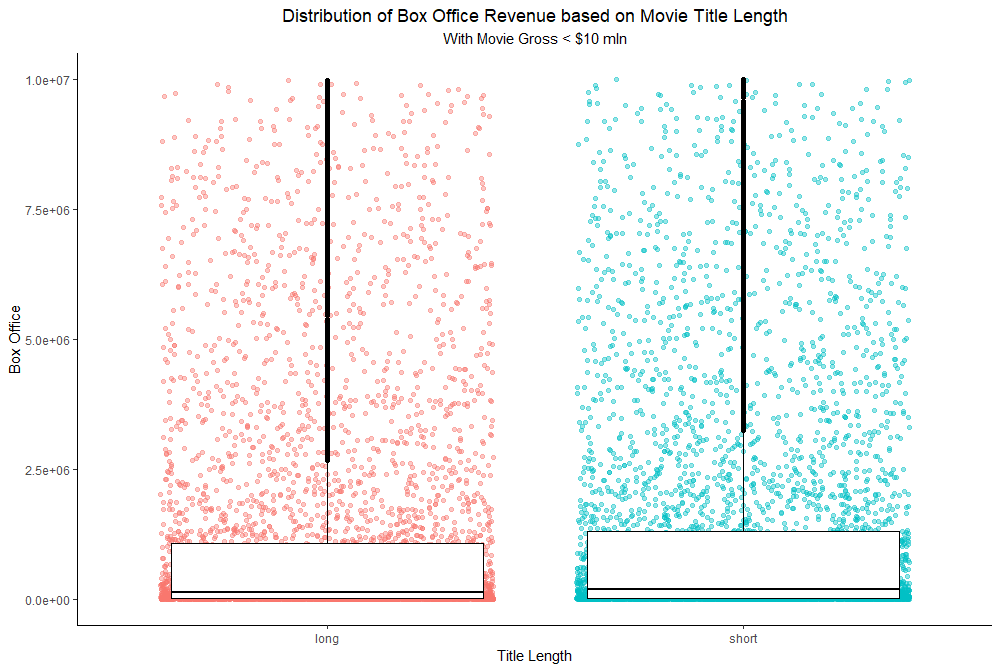
mutate(character\_length = nchar(title)) %>%

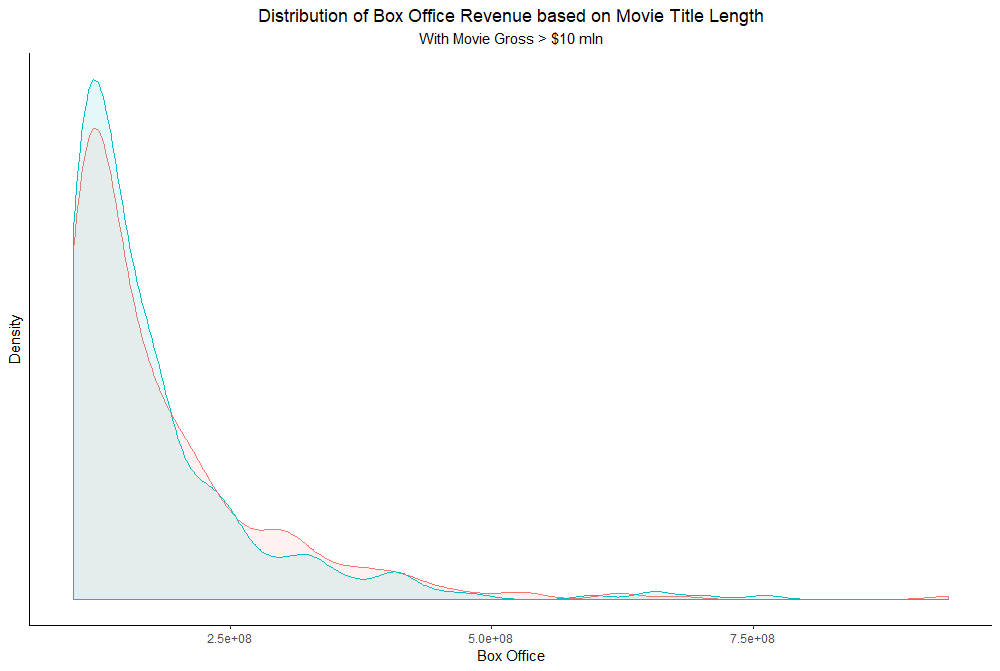
mutate(length = ifelse((character\_length <= 14), yes = "short", no = "long")) %>%

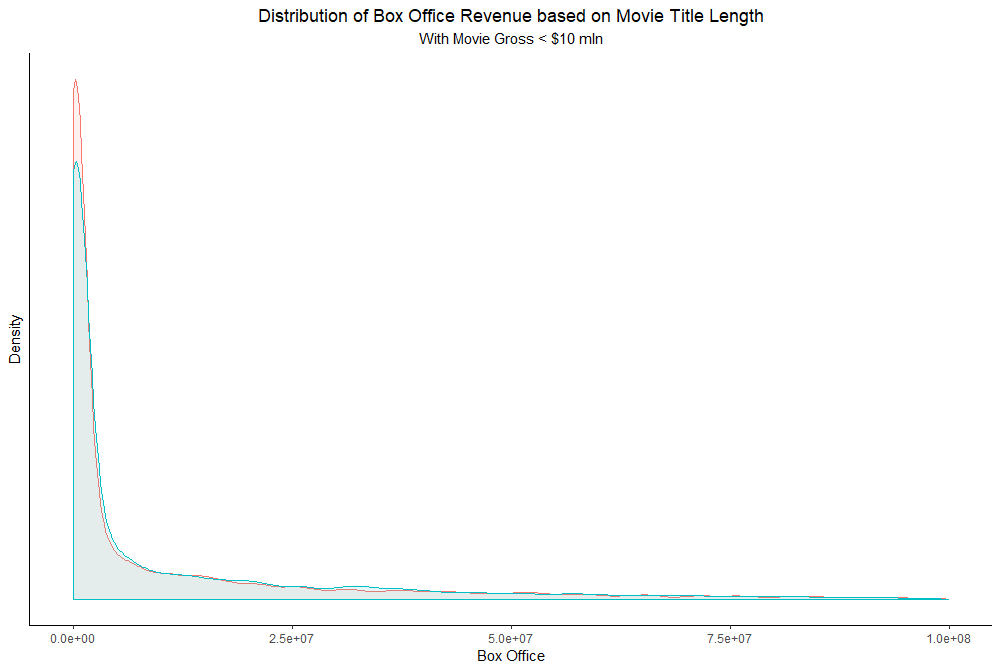
group\_by(length)

Next, using jitter and boxplot, I tried to plot the full dataset, which turned out to be a horrible idea for visual representation. The outliers have a huge impact on the data, so I decided to break the data apart into 2 additional categories: serious contenders (> $10 mln.) and never-heard-of’s (< $10 mln.). The boxplots show almost identical distribution between movies with short and long titles, with medians falling within $450 million of each other (which is kind of remarkable, considering that totals for both groups are in hundreds of billions of dollars). To get a different view of the distribution, I plotted density graphs which only confirmed a nearly identical distribution.





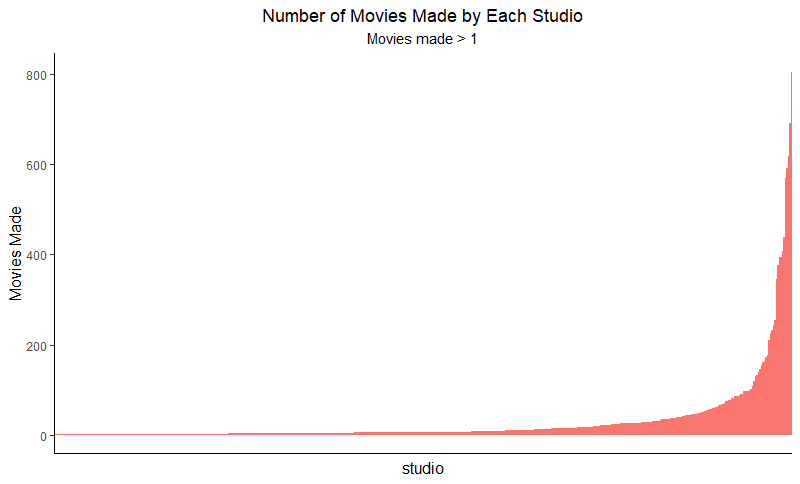
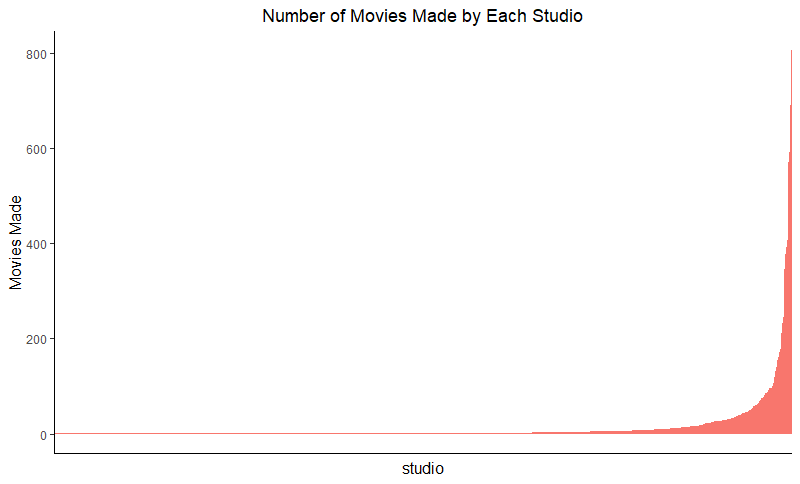




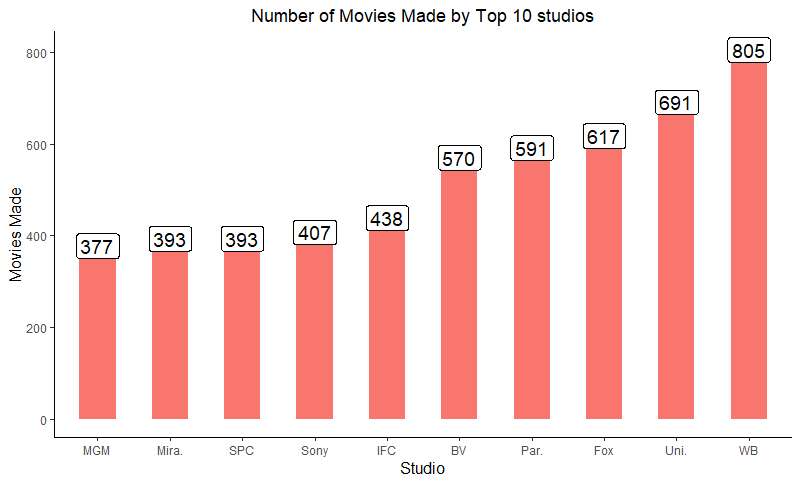
# Studios

Shifting from movies to studios, first I wanted to visualize historic distribution of movies made per studio. Plotting the whole dataset indicates that most of the studios have very few films made (95% of studios made <= 100 films). In hopes of producing a better histogram, I excluded all the studios that made only 1 film, and although there was a shift in distribution, it was still insignificant in terms of gaining insights. The bottom line – most of the studios out there have produced less than 10 films, and only very few have hundreds of titles under their belt.

IV. Distribution of Movies Made per Studio



The following are the giants of the industry with their total movies made since the inception of the industry. While most of them are familiar to the general public, others may be not so much. IFC is a NY based company owned by AMC; SPC is Sony’s American subsidiary that focuses solely on documentaries, independent and art films; Disney’s Buena Vista (BV) has been described as the “international distribution division of the Walt Disney Company”, and Metro-Goldwyn-Mayer (MGM), known for its iconic intro, does more co-production and co-financing than solo production these days.



V. Studio Competition & Market Saturation

While knowing how many movies each studio made, a better visualization is required to understand how studios have stack up against each other throughout the century.

The following 3 graphs show industry saturation and dominance at different eras.

**Historical Studio Density:**

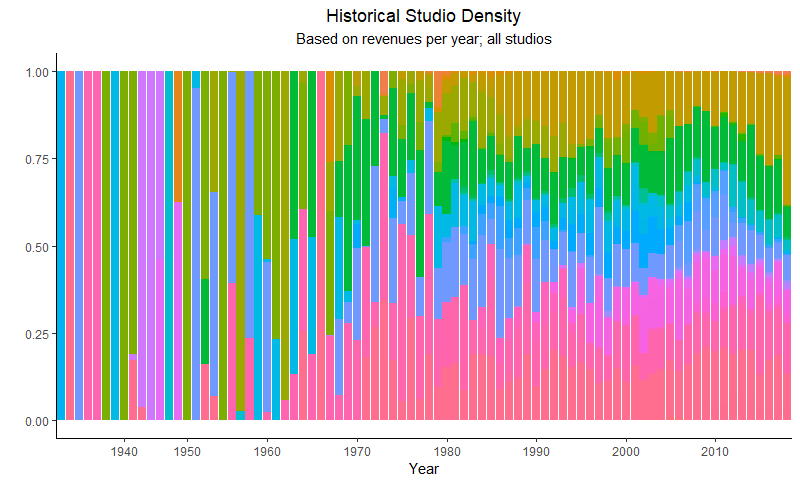
* No filters applied
* Too many studios to be represented, therefore most are lost during plotting, and legend is removed
* Serves as an abstract representation of how certain studios go from eras of dominance to almost complete disappearance, and how others can stay relevant over time (sign of a fast-paced industry).

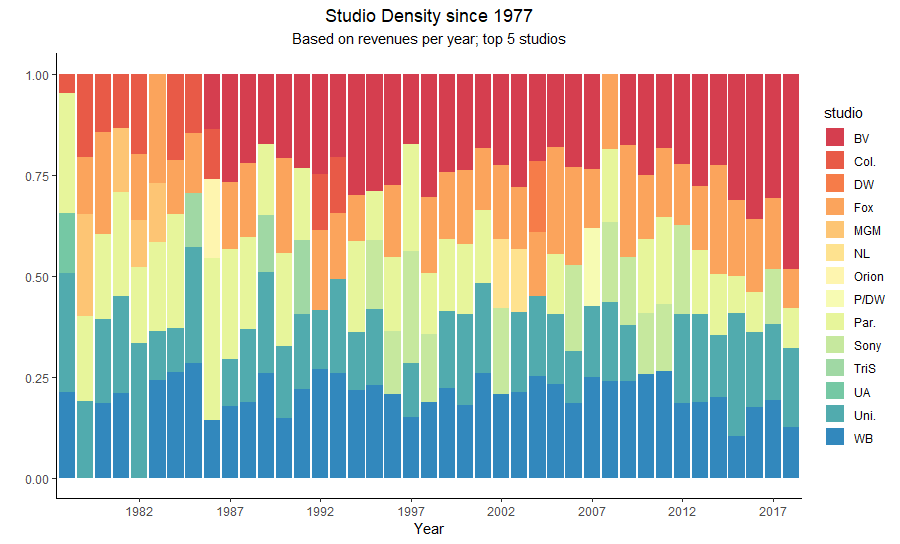
**Studio Density since 1977:**

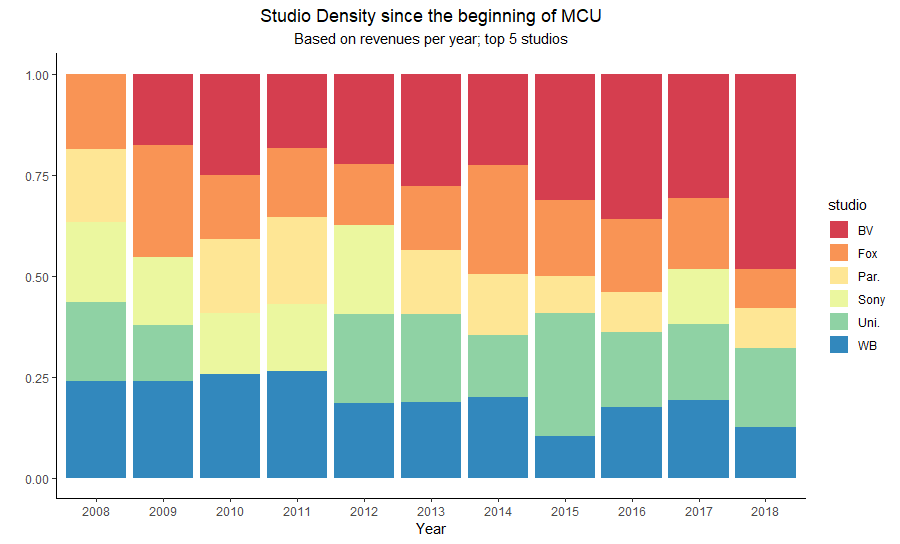
* Filtered from 1977 (Release of Star Wars and beginning of the golden era of cinema)
* Top 5 studios **per year** are shown, 14 total overall.
* Long term dominance by Buena Vista, Fox, Universal, and Warner Brothers, with occasional hits by Dreamworks, Sony and Paramount.
* Warner Brothers has been in Top 5 since 1983

**Studio Density since the beginning of MCU:**

* Filtered since 2008 (Iron Man), the beginning of new superhero era in cinema
* 6 Top studios selected based on their long-term performance and total box office, top 5 are shown each year.
* Disney’s BV became the market leader since the ramp of MCU; combined with the recent acquisition of Fox, Disney currently occupies more than 50% of the market (among top studios).
* Warner Brothers loses its share by failing to deliver the super hero hype with the DC comics universe; stays afloat thanks to *Rampage, Ready Player One, Lego Movie, Blade Runner* and *American Sniper.*
* Universal banks on *Jurassic World* franchise and *Despicable Me*, while Sony struggles to stay relevant.

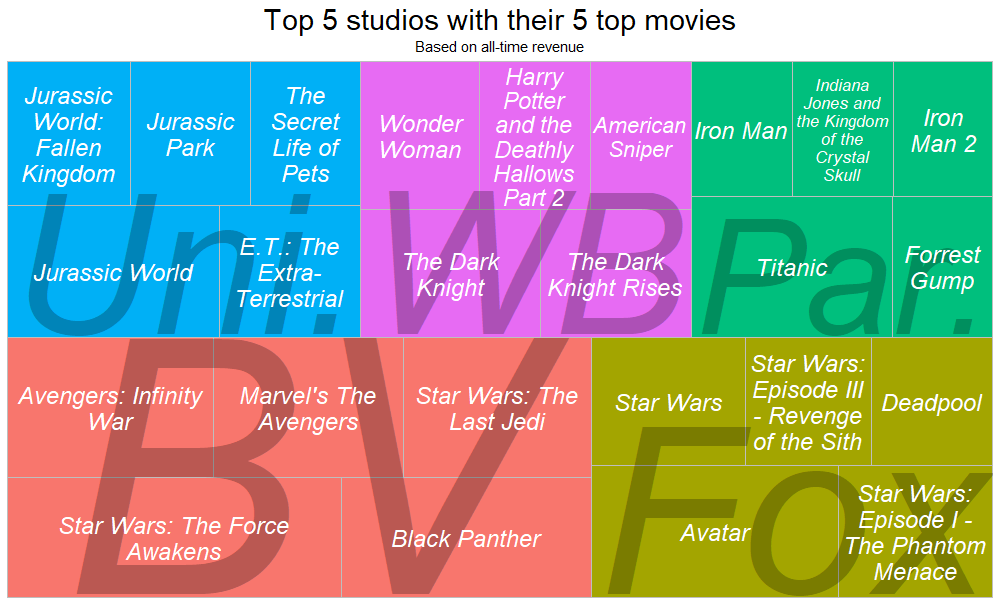


  
Graph 6



VI. Movies that Made Studios Money

These 5 studios are the titans of the industry, and the following treemap visualizes which movies played the biggest role in shaping their identities. The area of each square represents total box office revenue for that particular movie.

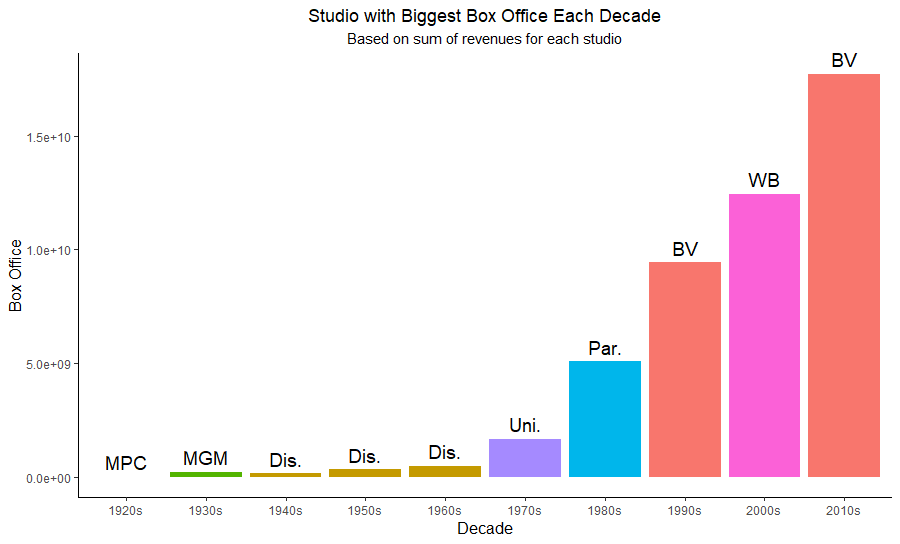


Out of 25 movies:

* 15 are originals (not sequels)
* Only 5 are not a part of a bigger franchise (Forrest Gump, Titanic, The Secret Life of Pets, E.T., American Sniper)
* 6 are a part of a reboot/soft reboot franchise
* 9 are superhero movies
* 19 were released after 2000

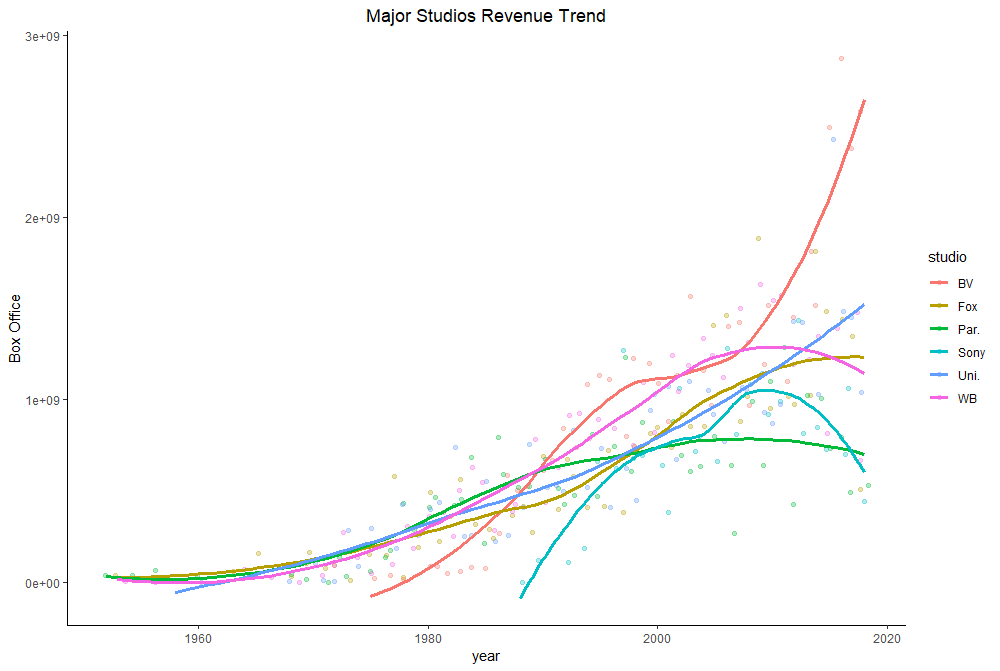
Additionally, only 2 out of 8 Star Wars movies are not represented, and all 5 of Buena Vista’s movies were released in the last 6 years.

VII. Which movie studio was the most successful at attracting people with their product each decade?



* Rise of Disney animation in the mid 20th century gifted us with such wonders as *Bambi*, *Pinocchio*, *Cinderella*, *Peter Pan*, *Alice in Wonderland*, *Lady and the Tramp*, *Sleeping Beauty*, *Marry Poppins* etc. Although the revenues seem insignificant in comparison to today’s numbers, the studio stayed above the pack for 3 decades in a row.
* Paramount’s 1980’s were marked with the beginning of the *Indiana Jones* Franchise, *Airplane*, and *Coming to America*.
* 90’s were a transformational period for Disney’s Buena Vista. With many new partnerships and release deals, titles like *The Lion King*, *Alladin*, *Mulan*, *Toy Story*, and *Tarzan* drove the studio’s profits for an entire decade.
* Warner Brothers made a bet on movies with dark aesthetic and protagonists with well developed backstories, giving rise to titles like *The Dark Knight*, *Watchmen*, *The Departed*, *300*, and *The Matrix*, *Harry Potter*, *Ocean’s Eleven* and *300* franchises.
* 2010 is the decade of nostalgia, and Disney has made a successful bet on audience craving more superhero movies in their daily lives. The purchase of rights to the *Star Wars* IP has allowed Disney to transfer their expertise from the MCU into the LucasArts projects, bringing them in total of over 2 billion dollars in worldwide box office revenue for The Force Awakens.

And this graph depicts the underlying trends behind these studios’ success. “Loess” smoothing was picked to better emphasize the trends of box office revenues throughout the years, and where the studios might potentially end up in the future.



VIII. Studios with highest ROI

While we can’t calculate actual ROI (the data doesn’t supply us with the costs, and even those wouldn’t be accurate, considering how vague the marketing costs are in the movie industry), we can approximate which studios are better at getting the most buck for their effort. To find that average “ROI”, we need to divide each studio’s total revenue by the number of movies they produced.

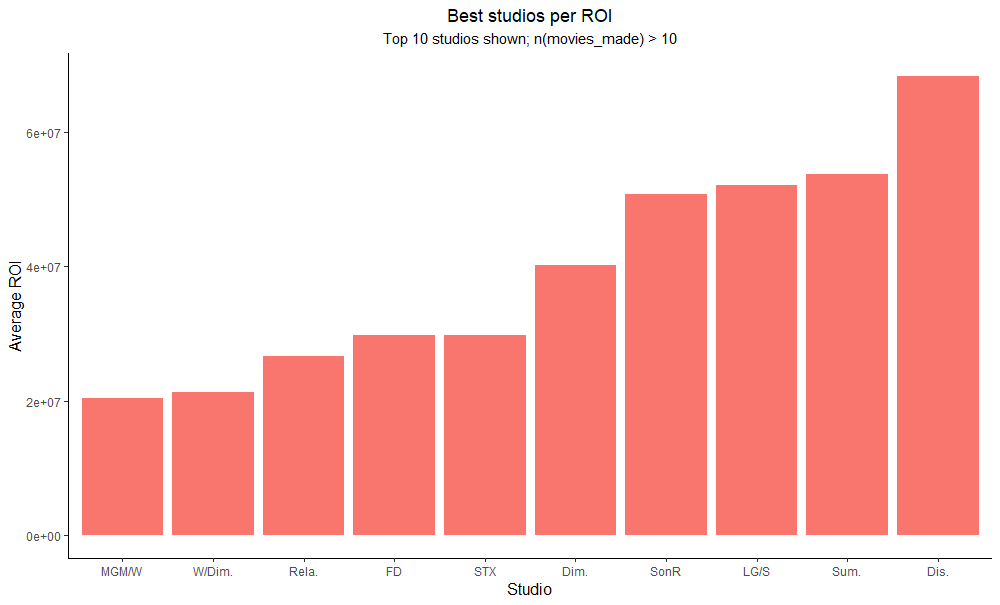
boxoffice %>%

group\_by(studio) %>%

summarise(total\_revenue = sum(lifetime\_gross), number\_of\_movies = n()) %>%

mutate(avg\_return\_on\_investment = total\_revenue / number\_of\_movies)

And to get more valuable insights, I decided to filter number of movies made > 10, since a studio that has produced at least 10 movies is not really who we want to observe, plus, with this formula, there is an unfair advantage for studios who only made 1-2 movies but were lucky enough for those to become big hits.



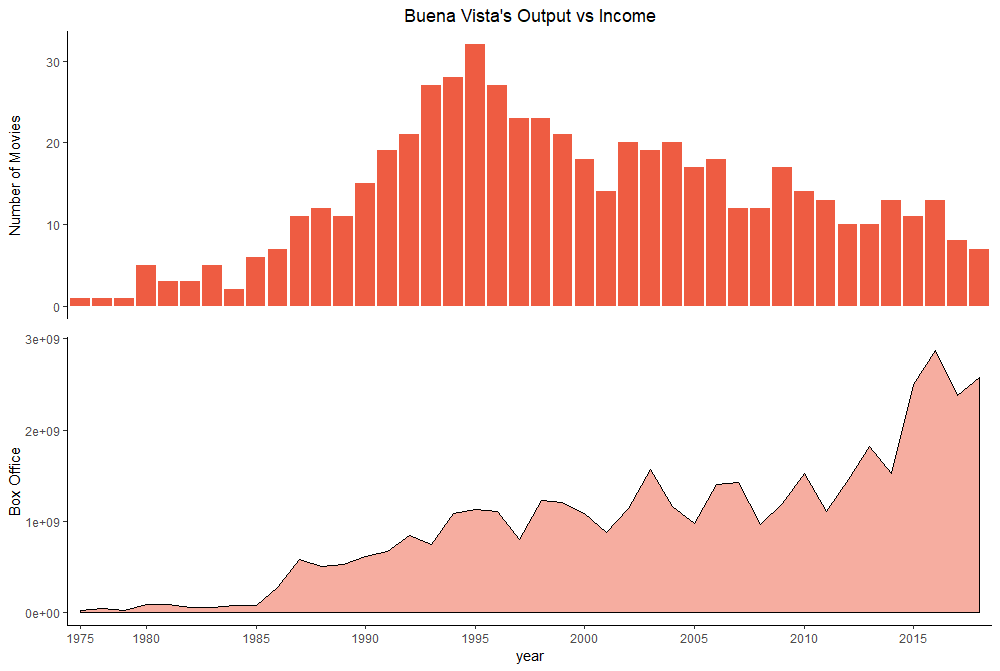
In my understanding, this is our best visualization of quality over quantity, and Disney manages to succeed at both. However, this bar chart also gives us a good idea how medium-sized studios can stay afloat – by appealing to more specific audiences, they can survive on smaller box office revenues (and keeping their costs down).

Buena Vista

XI. Overview of Output vs Revenue

Let’s take a look at BV – a Disney’s “subsidiary” that has become the biggest and the most studio in the world.

The following graph shows how their product output compares to their box office success. Interestingly, while the number of movies that Disney releases has declined significantly since the 90’s, their revenues are highest they’ve ever been. This explains their position in the “ROI” chart. While we don’t know the actual costs that Disney spares on movies these days vs. their production costs in the 90’s, it’s fair to say that they have a good grasp on what will appeal to the widest audience.



X. Disney vs Industry’s Average

In the next chart, I wanted to see a set of specifically-picked Buena Vista’s movies would stack up against the industry’s average. By making a (probably false) assumption that our data is normally distributed, I calculated the z score for every BV’s movie, arranged them from worst to best, and then picked random movies at specific intervals:

boxoffice2$gross\_z <- round((boxoffice2$lifetime\_gross - mean(boxoffice2$lifetime\_gross))/sd(boxoffice2$lifetime\_gross), 2)

boxoffice2$gross\_type <- ifelse(boxoffice2$gross\_z < 0, "below", "above")

boxoffice2 <- boxoffice2[order(boxoffice2$gross\_z),]

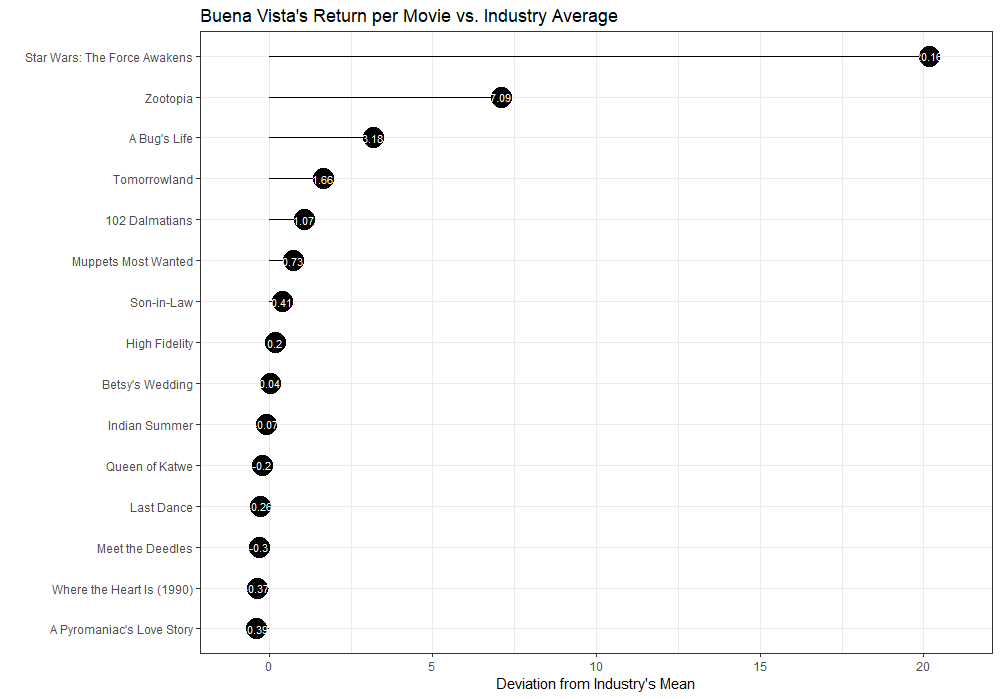
boxoffice2 <- filter(boxoffice2, studio == "BV")

boxoffice2$title <- factor(boxoffice2$title, levels = boxoffice2$title)

boxoffice2selected <-

boxoffice2[c(1,25,50,75,100,150,200,250,300,350,400,450,500,550,570)]

In the chart, 0 on x axis represent industry’s average box office return.



*A Pyromaniac’s Love Story*, Disney’s project with the worst z score, is only 0.39 standard deviations away from the industry’s mean, while the most successful movie that Disney has ever put out, is 20 standard deviations away from the mean. While we can’t measure how much effect does Disney’s brand name carries when it comes to appealing to the movie-goers, it seems that even at their worst, they still manage to be super close to what the rest of the industry considers to be an “average” box office. In hindsight, filtering this graph to the more recent data (like > 1990) would give us a higher industry average, but because of the Disney’s success in the 2010’s, I don’t expect this graph to look significantly different.

Conclusion, Limitations and Comments

The goal of this EDA was to allow me to practice newly learned R and specific packages, like ggplot2 and dplyr. The dataset was picked because of my love for cinema and the simplicity that it offered when it came to working with variables. While I would love to work with much more sophisticated movie datasets, that would allow me to investigate relationships between movie descriptors and the box office success of movies, I found a few projects online that have already met that objective, and I wouldn’t be able to (a) match the quality of their work, (b) come with new questions to ask given the depth and quality of their analyses.

In terms of accuracy of certain graphs and queries, outliers definitely had an impact on proper representation. The problem is that the data ranges from periods when the movie industry was virtually non-existent to the present day where multi-billion-dollar titles are becoming a norm. Separating the analysis by decades kind of negates this effect, but it is hard to distinguish what is a good criterion for something be considered an outlier, and what is the new average.