Analyzing Movie Franchises

Vedang Mehta

Advisor: Prof. Steven Buyske

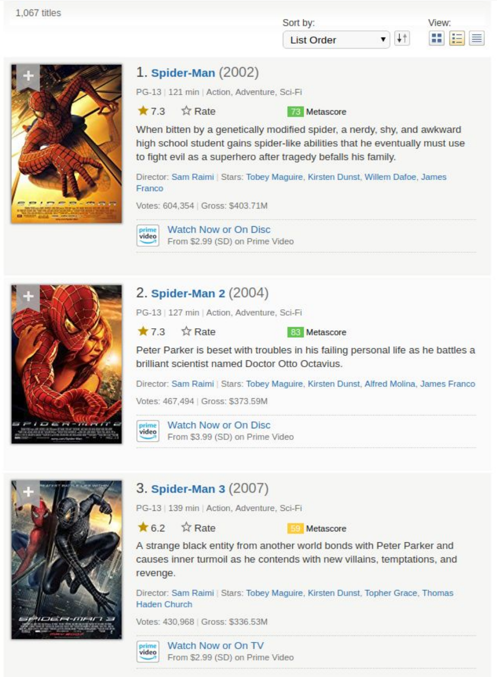
Objective

Getting data about movie franchises by scraping IMDb website and using OMDb API, and trying to get interesting insights.

Outline

- Get list of movie sequels from IMDb.
- Gather data about these movies using OMDb API.
- Figure out which movies belong to the same franchise.
- Tidying up the dataset.
- Find interesting insights.

Getting the Dataset



1,067 titles

Sort by: List Order

View:

1. Spider-Man (2002)

PG-13 | 121 min | Action, Adventure, Sci-Fi

★ 7.3 ☆ Rate 73 Metascore

When bitten by a genetically modified spider, a nerdy, shy, and awkward high school student gains spider-like abilities that he eventually must use to fight evil as a superhero after tragedy befalls his family.

Director: Sam Raimi | Stars: Tobey Maguire, Kirsten Dunst, Willem Dafoe, James Franco

Votes: 604,354 | Gross: $403.71M

Watch Now or On Disc
From $2.99 (SD) on Prime Video

2. Spider-Man 2 (2004)

PG-13 | 127 min | Action, Adventure, Sci-Fi

★ 7.3 ☆ Rate 83 Metascore

Peter Parker is beset with troubles in his failing personal life as he battles a brilliant scientist named Doctor Otto Octavius.

Director: Sam Raimi | Stars: Tobey Maguire, Kirsten Dunst, Alfred Molina, James Franco

Votes: 467,494 | Gross: $373.59M

Watch Now or On Disc
From $3.99 (SD) on Prime Video

3. Spider-Man 3 (2007)

PG-13 | 139 min | Action, Adventure, Sci-Fi

★ 6.2 ☆ Rate 59 Metascore

A strange black entity from another world bonds with Peter Parker and causes inner turmoil as he contends with new villains, temptations, and revenge.

Director: Sam Raimi | Stars: Tobey Maguire, Kirsten Dunst, Topher Grace, Thomas Haden Church

Votes: 430,968 | Gross: $336.53M

Watch Now or On TV
From $2.99 (SD) on Prime Video

- Found a list of sequels on IMDb. This list had 1000+ movies.
- Movies were listed in order of franchise, but there was no indication of separate franchises.
- Had to find a way to identify what franchise a movie belongs to.

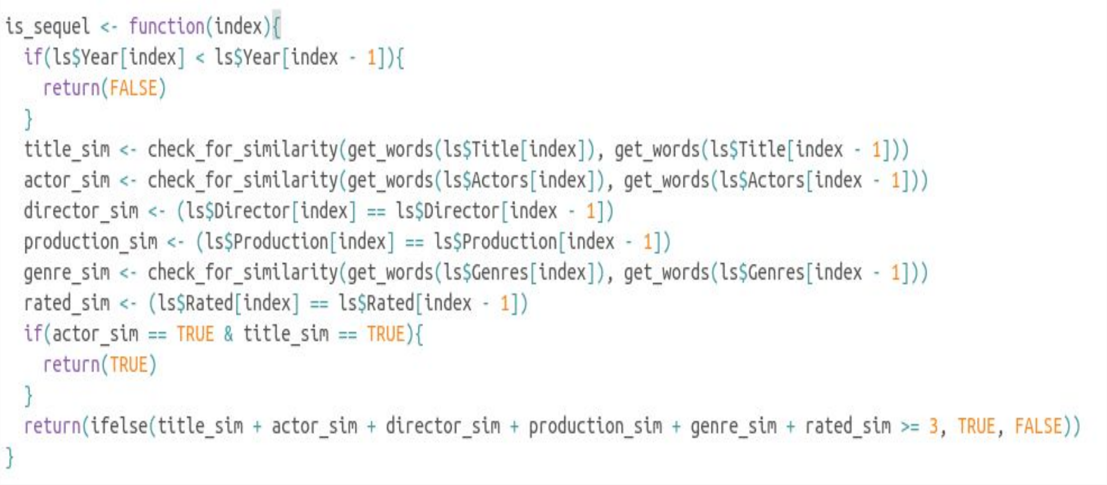
Tidying Up

- Attributes like actors were comma separated strings.
- Box-office collection was in currency ($100,000,000) format. Converted this column to numeric values with regular expression.
- Awards column was in "Won 2 Oscars. Another 11 wins & 20 nominations." format. Wrote a regular expression to get the number of Oscars.
- OMDb had a lot of missing values for box-office collection, which I had to scrape from IMDb.

Getting More Data with OMDb API

- Once I had the list of movie titles from IMDb, I could collect more information using OMDb API.
- OMDb API provided information such as movie title, IMDb rating, box-office collection, actors, director, producer, runtime etc.
- Data was collected in JSON format from OMDb and I stored it to a CSV file so that it becomes easier to process in the future.

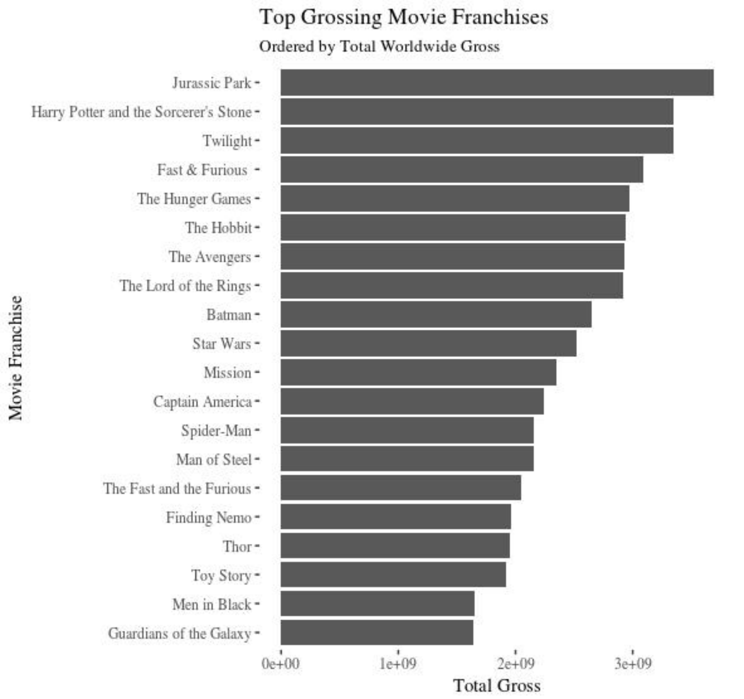
Identifying Movie Franchise

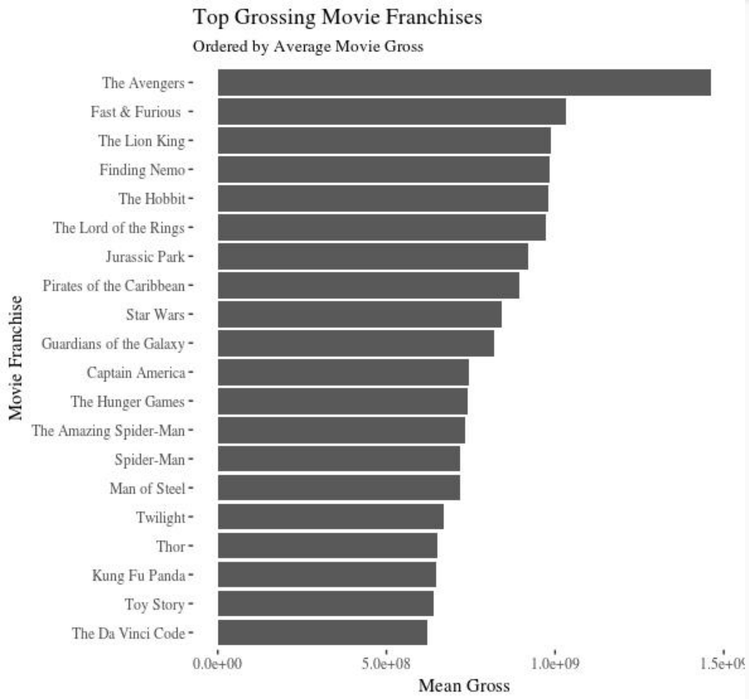


```
is_sequel <- function(index){
  if(ls$Year[index] < ls$Year[index - 1]){
    return(FALSE)
  }
  title_sim <- check_for_similarity(get_words(ls$Title[index]), get_words(ls$Title[index - 1]))
  actor_sim <- check_for_similarity(get_words(ls$Actors[index]), get_words(ls$Actors[index - 1]))
  director_sim <- (ls$Director[index] == ls$Director[index - 1])
  production_sim <- (ls$Production[index] == ls$Production[index - 1])
  genre_sim <- check_for_similarity(get_words(ls$Genres[index]), get_words(ls$Genres[index - 1]))
  rated_sim <- (ls$Rated[index] == ls$Rated[index - 1])
  if(actor_sim == TRUE & title_sim == TRUE){
    return(TRUE)
  }
  return(ifelse(title_sim + actor_sim + director_sim + production_sim + genre_sim + rated_sim >= 3, TRUE, FALSE))
}
```

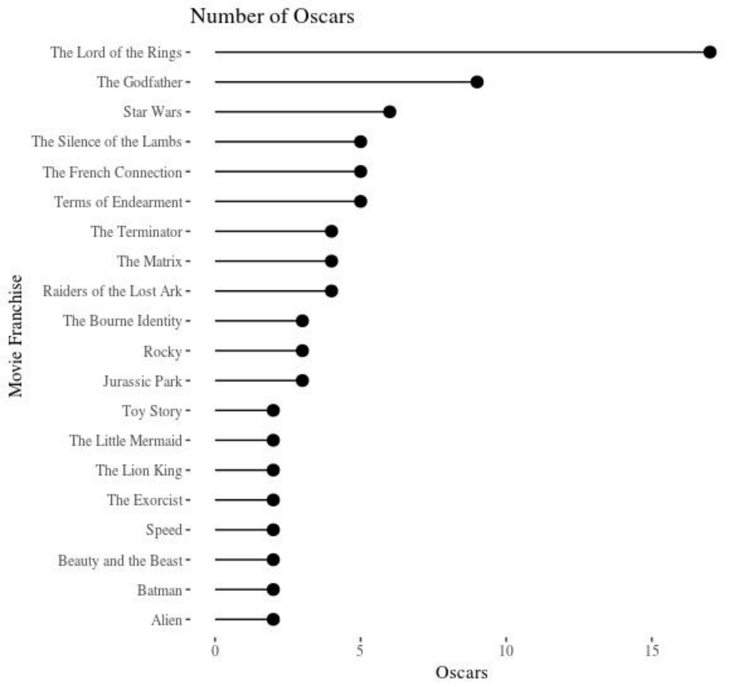
- No need to perform any complex analysis.
- Movies were already ordered by their franchise.
- Comparing basic attributes of two consecutive movies gave satisfactory results.

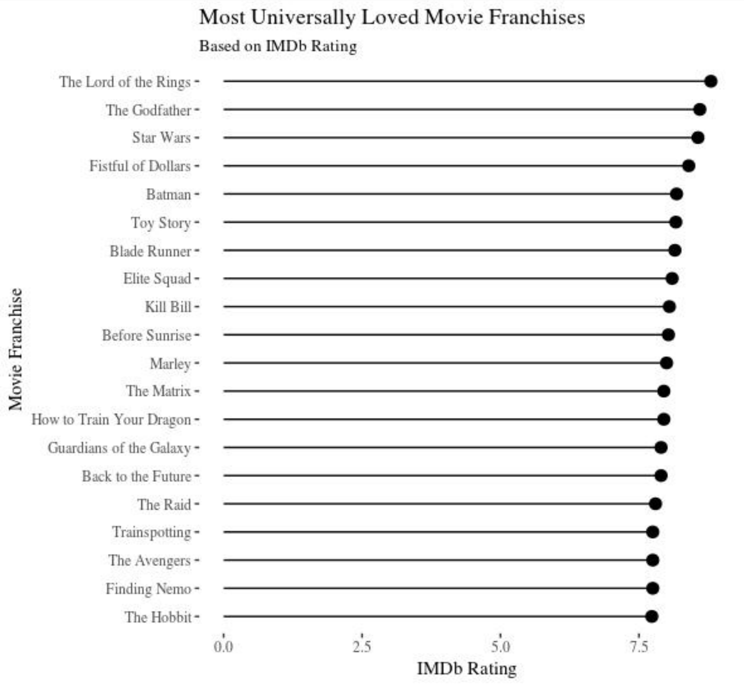
Top Grossing Franchises





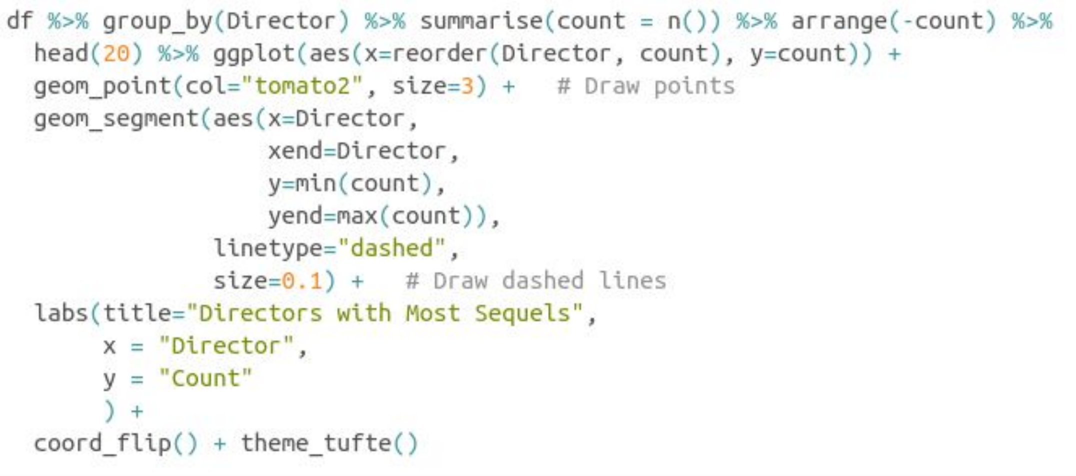
Most Universally Loved Movie Franchises



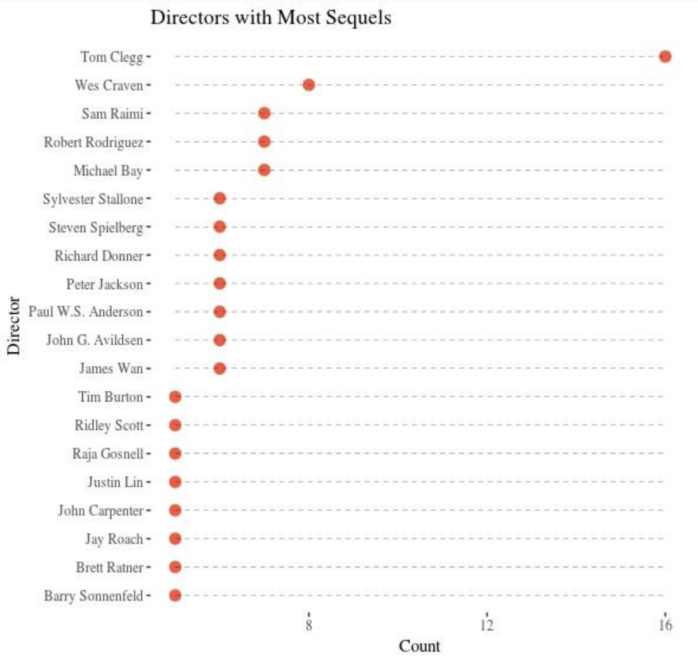


Directors with Most Sequels

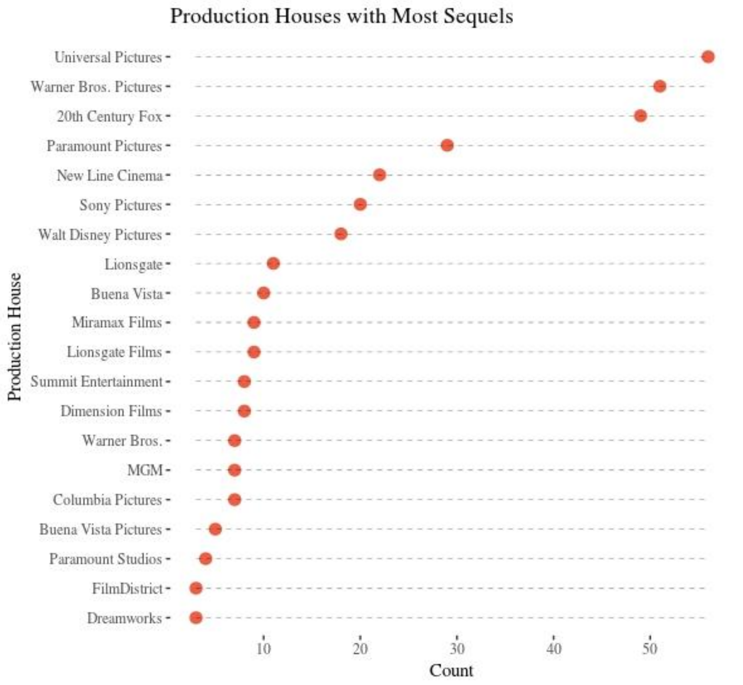
Code for Dot Plot



```
df %>% group_by(Director) %>% summarise(count = n()) %>% arrange(-count) %>%
  head(20) %>% ggplot(aes(x=reorder(Director, count), y=count)) +
  geom_point(col="tomato2", size=3) +   # Draw points
  geom_segment(aes(x=Director,
                   xend=Director,
                   y=min(count),
                   yend=max(count)),
               linetype="dashed",
               size=0.1) +   # Draw dashed lines
  labs(title="Directors with Most Sequels",
       x = "Director",
       y = "Count"
       ) +
  coord_flip() + theme_tufte()
```

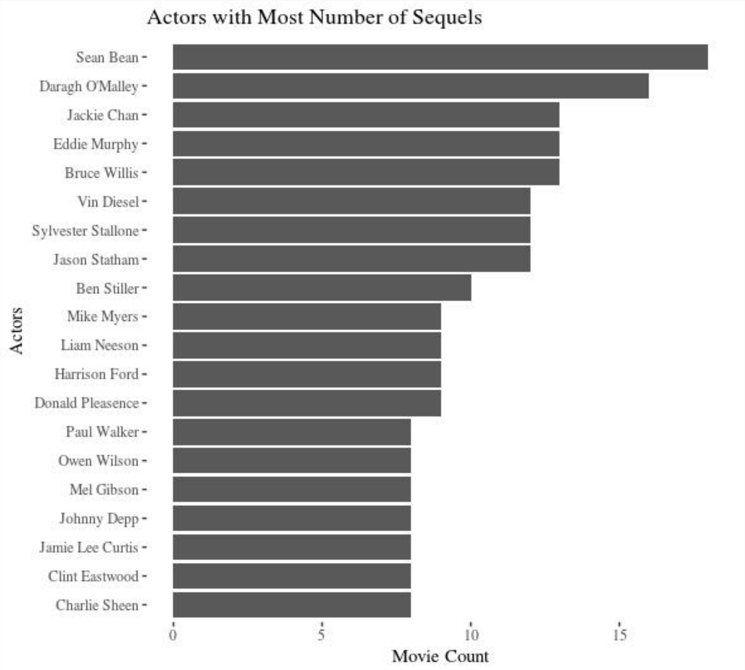


Production Houses with Most Sequels



Top three production houses produced significantly higher number of sequels than the rest.

Actors with Most Sequels



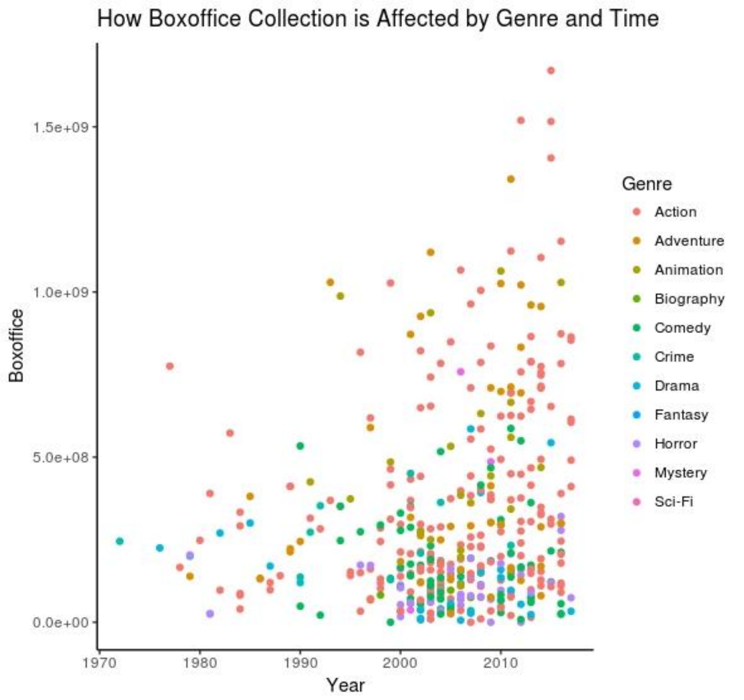
As said before, actors were separated by commas. (example, "Seth Rogen, James Franco, Jonah Hill").

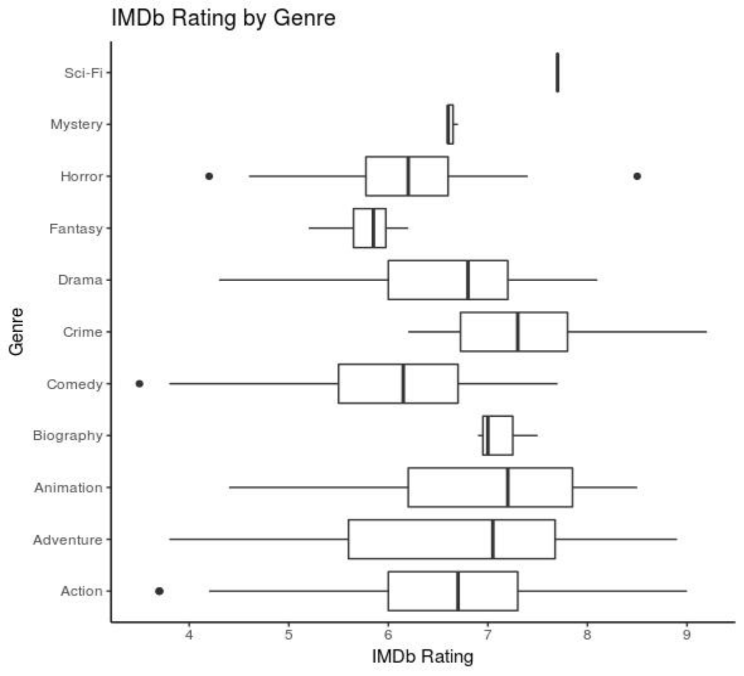
Wrote the following code to convert Actors column to a single list of actors -

```
strsplit(df$Actors, ", ") %>% rbind() %>%
apply(MARGIN = 1, unlist)
```

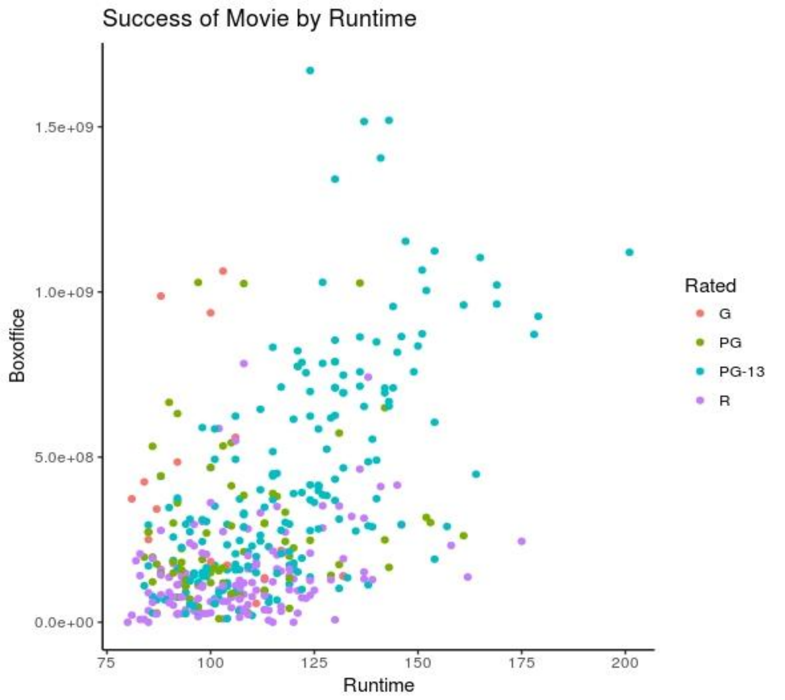
Performed count on this list and plotted a bar plot.

Does Genre Affect the Success of the Movie?





Effect of Runtime on Box-office Success



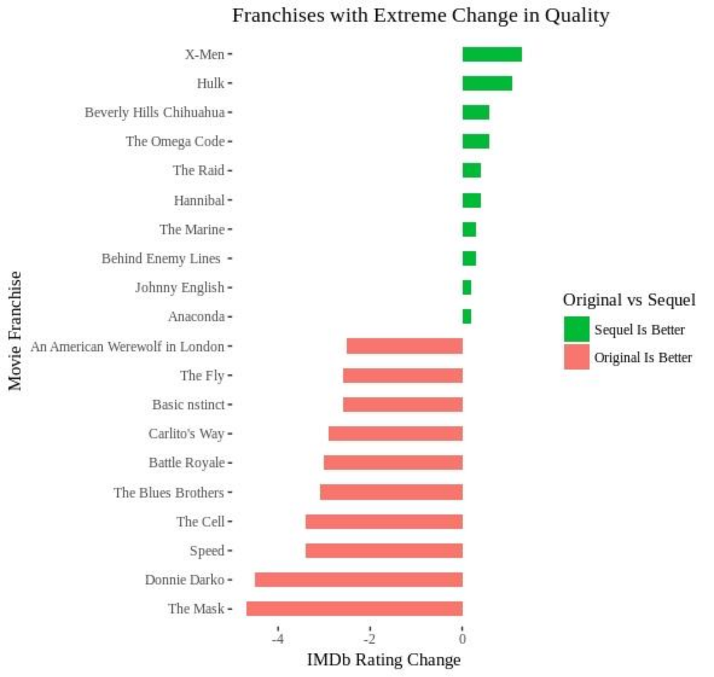
- The correlation between runtime and box-office collection is not very strong.
- Even though the correlation isn't very strong, longer movies tend to earn a bit more.
- Box-office performance correlates more strongly to how the movie is rated.

Future Scope

- Visualize what franchises are getting better and what franchises are getting worse.*
- Show comparison with inflation adjusted box-office collection.
- How often do sequels perform better than the original movies?*

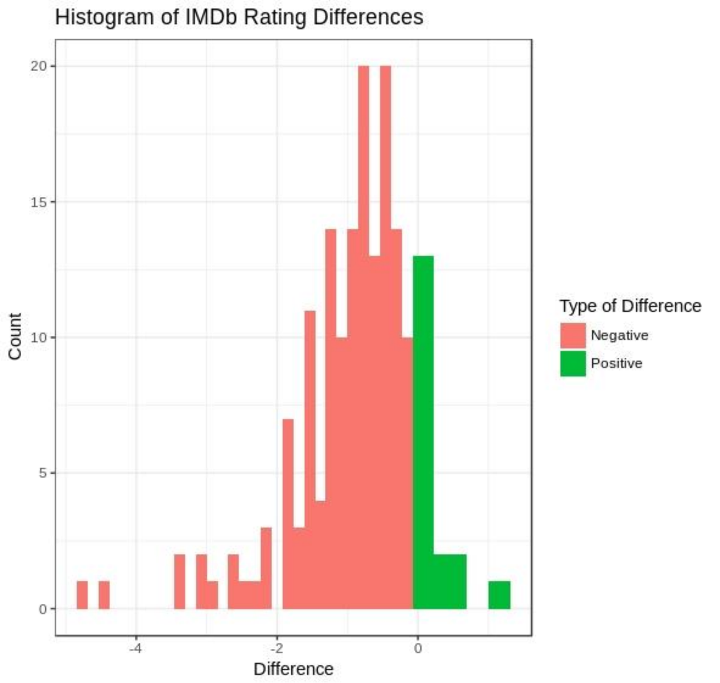
*Added after the in-class presentation. See the next two slides.

Franchises that Got Significantly Better or Significantly Worse



- Only considering franchises with two movies.
- There were 188 such franchises in the dataset.
- Displaying 10 movies with highest positive difference and 10 movies with highest negative difference in IMDb ratings.
- Used diverging bars to visualize positive/negative change in rating.

How Often Do Sequels Perform Better than the Original?



- Created a histogram of IMDb rating differences.
- Only 34/188 franchises in the dataset with the sequel having higher rating than the original.
- Magnitude of rating difference is also higher on the negative side than the positive.

References

1. OMDb API - http://www.omdbapi.com/
2. IMDb to collect the list of movie sequels - https://www.imdb.com/list/ls003495084/
3. Gallery of ggplot2 visualizations - http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Thank You!