Data Wrangling Final Project - Popular Movies

Eunbin Ko

2019 5 1

Introduction

These days, people can easily find out the evaluation of movies from various "movie rating" websites. However, not all the movie rating websites has the same result. In order to find out similarities of movies in various movie websites, I chose IMDb, Netflix, and Box-office to analyze. All the websites rate movies into various way, and among those availabilities, I decided on analyzing top 100 movies that have been specified as "popular" in IMDb and Netflix, then find out how these relate to the top 100 worldwide grosses, which shows in Box-Office website. For the rank in the IMDb chart, which reflects the popularity, is determined by IMDb Users. However, for the rank of popular movies in Netflix, the website did not specify how the popularity is being ranked, instead it gives users to see the movies that is "sort by popular". I assume the popularity of movies in Netflix is based upon Netflix Users.

Datasets

1.IMDb

First, go to the IMDb's top 100s most popular movies from link: "https://www.imdb.com/chart/moviemeter?ref_=nv_mv_mpm" Get the data by scraping the webpage. Below is the example of scraped data before the cleaning.

```
##
## 1 NA
## 2 NA
## 3 NA
## 4 NA
## 5 NA
## 6 NA
##                                                                            Rank & Title
## 1                        Avengers: Endgame\n        (2019)\n            1\n(no change)
## 2                       Avengers: Infinity War\n        (2018)\n            3\n(\n\n1)
## 3 Extremely Wicked, Shockingly Evil, and Vile\n        (2019)\n            4\n(\n\n53)
## 4                               Captain Marvel\n        (2019)\n            5\n(\n\n2)
## 5                                   Long Shot\n        (2019)\n            6\n(\n\n25)
## 6                   Spider-Man: Far from Home\n        (2019)\n            7\n(\n\n13)
##   IMDb Rating
## 1         8.9
## 2         8.5
## 3         6.7
## 4         7.1
## 5         7.2
## 6          NA
##
## 1 12345678910\n        \n        \n            \n            NOT YET RELEASED\n             \n
## 2 12345678910\n        \n        \n            \n            NOT YET RELEASED\n             \n
## 3 12345678910\n         \n        \n            \n            NOT YET RELEASED\n             \n
## 4 12345678910\n         \n        \n            \n            NOT YET RELEASED\n             \n
## 5 12345678910\n         \n        \n            \n            NOT YET RELEASED\n             \n
## 6 12345678910\n         \n        \n            \n            NOT YET RELEASED\n             \n
##   
## 1 NA
## 2 NA
## 3 NA
## 4 NA
## 5 NA
## 6 NA
```

There are movies' titles with their release year sorted by rank, and the ratings of the movies. Clean the data by having three columns "Title", "Year", and "Rating" in the order of rank. While cleaning the data, I had to remove unnecessary symbols for variable "Year". For those rating that are not shown in the website, it will be listed as NA in R. Now the data for IMDb is cleaned and below is the output of first few lines of cleaned IMDb data.

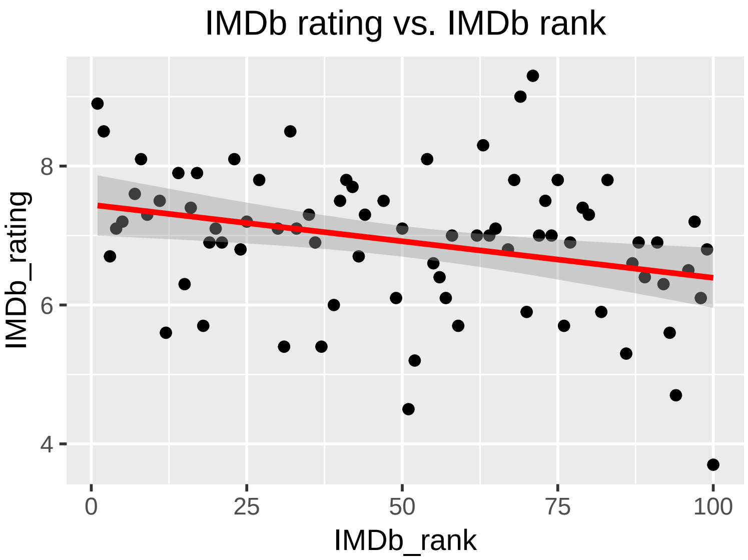
```
## Warning: Expected 2 pieces. Additional pieces discarded in 100 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
##                                        Title Year IMDb_rating IMDb_rank
## 1                          Avengers: Endgame 2019         8.9         1
## 2                     Avengers: Infinity War 2018         8.5         2
## 3 Extremely Wicked, Shockingly Evil, and Vile 2019         6.7         3
## 4                             Captain Marvel 2019         7.1         4
## 5                                  Long Shot 2019         7.2         5
## 6                  Spider-Man: Far from Home 2019          NA         6
```

Using IMDb dataset, there shows every movie's rating corresponding to its rank of the popularity. Let's fit a linear regression line to these datasets to see if there exists a linear relationship between the rating of the movie and the rank of the popularity. Below is the linear plot of the IMDb rating vs. IMDb rank.

```
## Warning: Removed 22 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 22 rows containing missing values (geom_point).
```



This plot clearly shows there is no linear relationship between rating and rank based on IMDb dataset. Using this result, we can conclude that rating of movies does not affect the popularity.

2.Box Office

Then go to Box-office website that gives 100 movies sorted in order of worldwide grosses: “https://www.boxofficemojo.com/alltime/world/”. When scraping this data, the first row which contains column names needs to be adjusted. After cleaning the data, select only columns that will be needed in this project (you can select more columns if necessary by changing “box_office %>% select(”column name“)” at the final line of “Box Office” section. For this project, I selected columns that contains “Box_office_rank”, “Title”, “Worldwide_gross”, and “Year”. For the csv file that contains box-office data, I have saved all the columns for further analysis in the future. The example of the cleaned dataset that is needed is shown below.

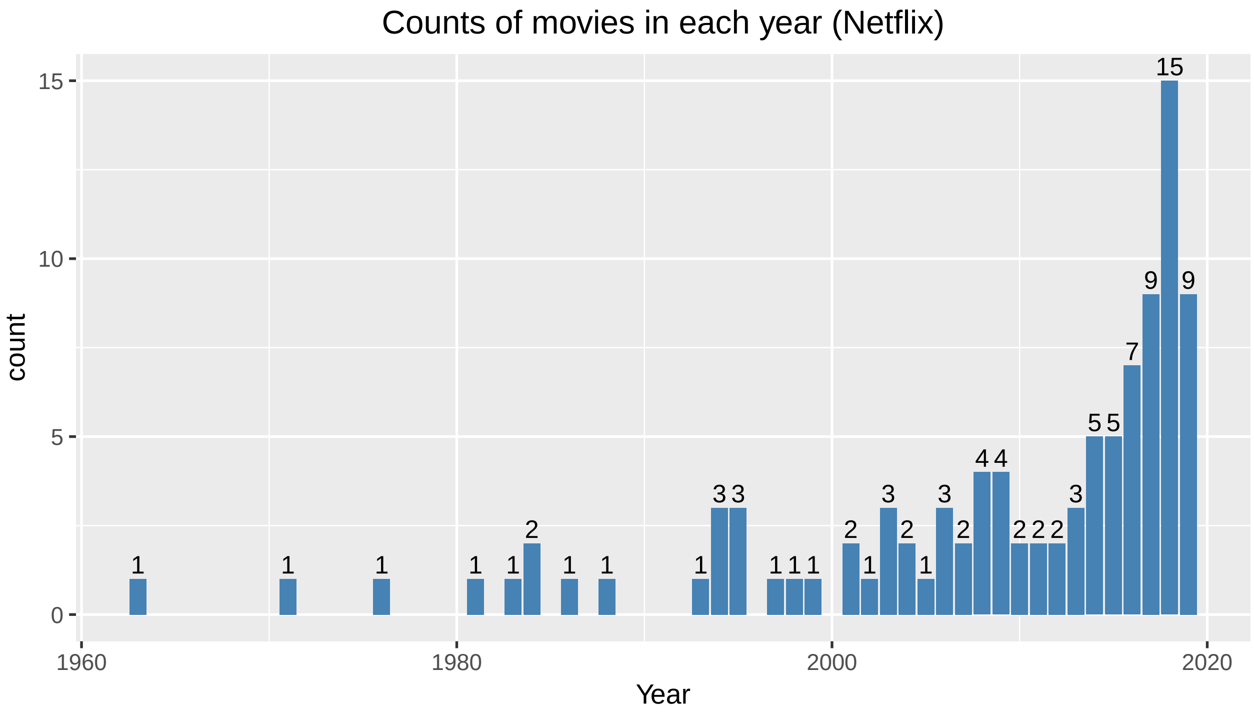
```
##   Box_office_rank                        Title Worldwide_gross Year
## 1               1                       Avatar          2788.0 2009
## 2               2            Avengers: Endgame          2193.7 2019
## 3               3                      Titanic          2187.5 1997
## 4               4 Star Wars: The Force Awakens          2068.2 2015
## 5               5       Avengers: Infinity War          2048.4 2018
## 6               6               Jurassic World          1671.7 2015
```

3.Netflix

Finally for Netflix data, I got them from "https://reelgood.com/movies/source/netflix?filter-sort=1" and "https://reelgood.com/movies/source/netflix?filter-sort=1&offset=50", which first link contains the first 50 of the dataset and the second link contains the second 50 of the dataset that has been sorted by popularity. The title for column needs to be fixed. This dataset contains two rating, one refers to the movie's suitability for certain audiences based on its content, and the other rating refers to how good the movie is. Therefore, in order to not to be confused with the rate that will be evaluated in this project, I will only consider the rating of "how good the movie" in here. Since this dataset is also sorted as popularity, I have added rank column to see the rank of each movie along with their ratings. Below is the example of Netflix dataset.

```
##                                         Title Year Netflix_rating
## 1                      Avengers: Infinity War 2018            8.5
## 2                               Black Panther 2018            7.3
## 3                              Thor: Ragnarok 2017            7.9
## 4 Extremely Wicked, Shockingly Evil and Vile 2019            6.7
## 5              Guardians of the Galaxy Vol. 2 2017            7.7
## 6                        Ant-Man and the Wasp 2018            7.1
##   Netflix_rank
## 1            1
## 2            2
## 3            3
## 4            4
## 5            5
## 6            6
```

Since these popular movies are from different released year (other data does too), let's see which year has the highest count of movies using bar plot.



According to the result of the bar plot, it can clearly see that movies in recent years are more likely to be favored by Netflix users. Among those years, year 2018 has the highest count of movies that are popular.

Comparisons

IMDb and Netflix

After collecting all the neccessary data from IMDb, Box-Office, and Netflix, now let's see the similarities between each datasets. First see the common movies in IMDb and Netflix by merging the two datasets from already cleaned datasets "imdb" and "netflix".

```
##                              Title Year IMDb_rating IMDb_rank
## 1          Ant-Man and the Wasp 2018         7.1        33
## 2        Avengers: Infinity War 2018         8.5         2
## 3                 Black Panther 2018         7.3        35
## 4  Guardians of the Galaxy Vol. 2 2017         7.7        42
## 5                 Someone Great 2019         6.1        49
## 6                      The Dirt 2019         7.0        74
## 7                The Highwaymen 2019         7.0        58
## 8               The Last Summer 2019         5.7        59
## 9              The Perfect Date 2019         5.9        82
## 10                  The Silence 2019         5.2        52
## 11               Thor: Ragnarok 2017         7.9        14
## 12              Triple Frontier 2019         6.5        96
##    Netflix_rating Netflix_rank
## 1             7.1            6
## 2             8.5            1
## 3             7.3            2
## 4             7.7            5
## 5             6.1           14
## 6             7.0           52
## 7             7.0            8
## 8             5.7            9
## 9             5.9           18
## 10            5.2           10
## 11            7.9            3
## 12            6.5           31
```

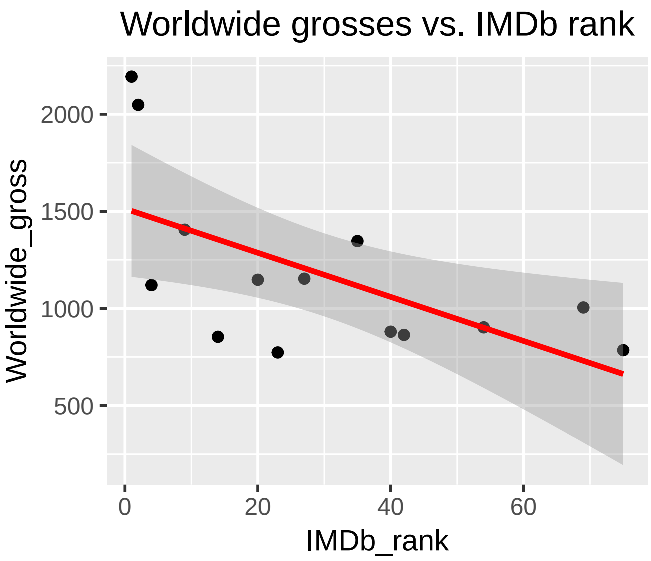
There are 11 movies in common that is shown in the top 100 popular movies in IMDb and Netflix website. We can see that the ratings of each movies stays the same but the rank is different. "Avengers: Infinity War" that has been released in 2018 shows the rank in first place for Netflix while IMDb has placed it in the second. Furthermore, among those 11 common movies, Netflix's rank contains all the first seven rank while IMDb's rank varies a lot. This show that althogh the ratings is same throughout the movies, each website's users' preference on movies varies.

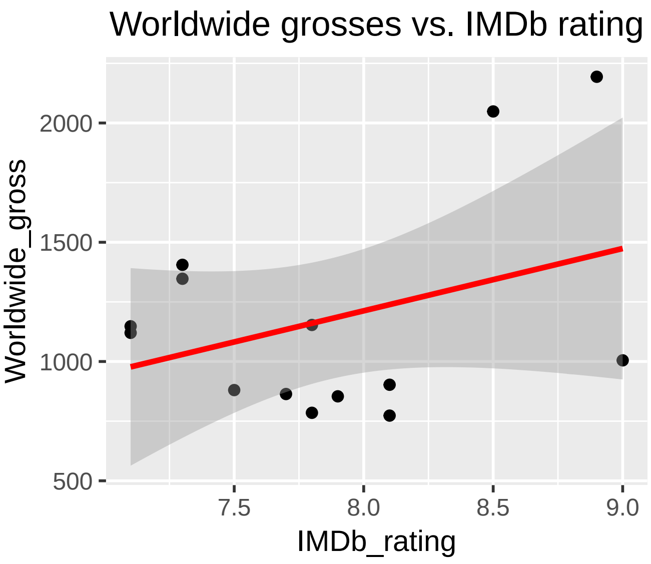
IMDb and Box-Office

Now let's see the relationship between top 100's worldswide grosses of movies in IMDb's top 100's popular movies. First, merge the two datasets of already cleaned "imbd" and "box_office" to see the common movies.

```
##                               Title Year IMDb_rating IMDb_rank
## 1                           Aquaman 2018         7.1        20
## 2           Avengers: Age of Ultron 2015         7.3         9
## 3                 Avengers: Endgame 2019         8.9         1
## 4            Avengers: Infinity War 2018         8.5         2
## 5                     Black Panther 2018         7.3        35
## 6                 Bohemian Rhapsody 2018         8.1        54
## 7        Captain America: Civil War 2016         7.8        27
## 8                   Captain Marvel 2019         7.1         4
## 9                       Deadpool 2 2018         7.8        75
## 10          Guardians of the Galaxy 2014         8.1        23
## 11 Guardians of the Galaxy Vol. 2 2017         7.7        42
## 12           Spider-Man: Homecoming 2017         7.5        40
## 13                  The Dark Knight 2008         9.0        69
## 14                   Thor: Ragnarok 2017         7.9        14
##    Box_office_rank Worldwide_gross
## 1               21          1147.7
## 2                9          1405.4
## 3                2          2193.7
## 4                5          2048.4
## 5               10          1346.9
## 6               54           902.6
## 7               20          1153.3
## 8               23          1120.2
## 9               87           785.0
## 10              91           773.3
## 11              67           863.8
## 12              59           880.2
## 13              39          1004.9
## 14              70           854.0
```

Above result shows that there are 14 movies are shown in both datasets and we can see the relationship between rank of the movie and the worldwide grosses. Let's fit a linear model between the worldwide gross and the rank of the movie, and also between worldwide gross and the rate of the movie.


Worldwide grosses vs. IMDb rank

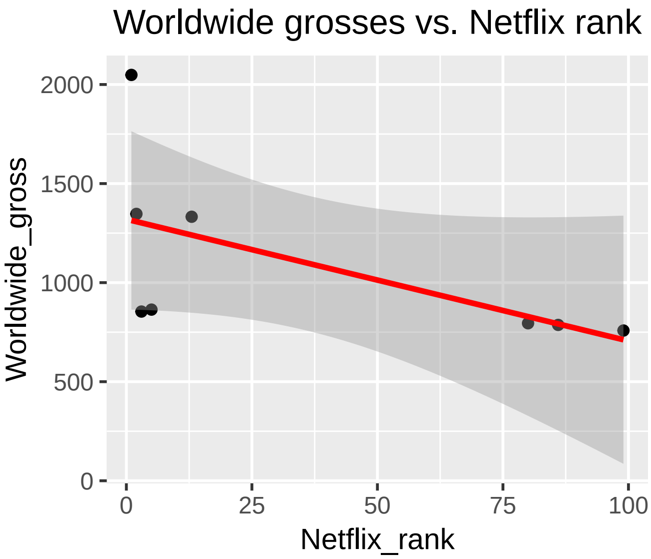

Worldwide grosses vs. IMDb rating

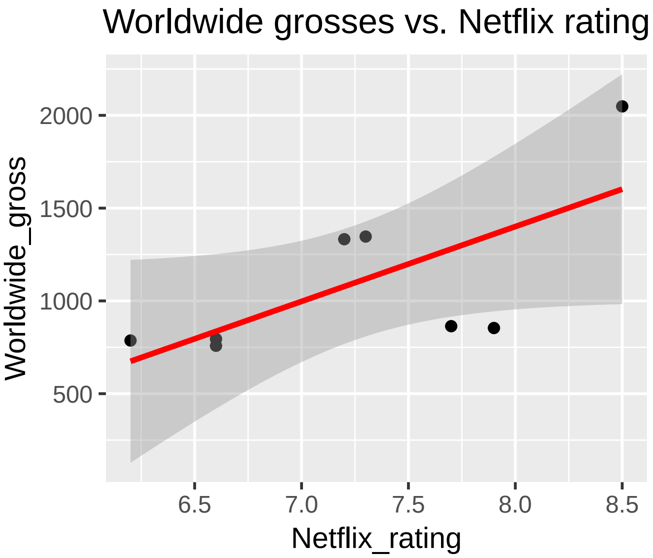
According to the linear regression result shown above, we can see the trend, which gives higer rank has better worldwide grosses. As for numerical result, we can conclude that the worldwide grosses decrease by 10.525 millions of dollars for every increase in rank in IMDb datasets (note that model only explains 33% of the variability of the response data around its mean). For worldwide grosses and IMDb rating plot, we can't conclude there is a linear relationship, and its Adjusted R-squared value is extremely small (6 percent), which indicates that the model is meaningless. This concludes that rating of the movie does not effect the worldwide grosses. However, we can see the broad trend of the model through the linear plot.

Netflix and Box-Office

Now let's see the relationship between Netflix and its worldwide grosses. Merge "netflix" and "box_office" to get the common movies in these datasets. There are 7 movies show in common. And also fit a linear plot of this merged dataset that gives the relationship between worldwide grosses and rank of Netflix. There are only half of the movies that is common with Netflix dataset and the Box-Office dataset compare to the IMDb dataset and the Box-Office dataset. This shows that IMDb website's popular movies contains more movies that have greater worldwide grosses.

```
##                                              Title Year Netflix_rating
## 1                           Avengers: Infinity War 2018            8.5
## 2                                    Black Panther 2018            7.3
## 3                   Guardians of the Galaxy Vol. 2 2017            7.7
## 4 Indiana Jones and the Kingdom of the Crystal Skull 2008            6.2
## 5   Pirates of the Caribbean: Dead Men Tell No Tales 2017            6.6
## 6                             Star Wars: The Last Jedi 2017            7.2
## 7                                    The Da Vinci Code 2006            6.6
## 8                                       Thor: Ragnarok 2017            7.9
##   Netflix_rank Box_office_rank Worldwide_gross
## 1            1               5          2048.4
## 2            2              10          1346.9
## 3            5              67           863.8
## 4           86              86           786.6
## 5           80              82           794.9
## 6           13              12          1332.5
## 7           99              94           758.2
## 8            3              70           854.0
```


Worldwide grosses vs. Netflix rank


Worldwide grosses vs. Netflix rating

According to the summary and plot shown above, there is 24 percent explaining the model that for every increase in rank, worldwide gross decreases by 6.145 millions of dollar. There are only 38 percent explaining the model that for every increase in rating(out of 10), worldwide grosses increases by 403.3 millions of dollar. We can see the increasing and decresing trend from the linear plot.

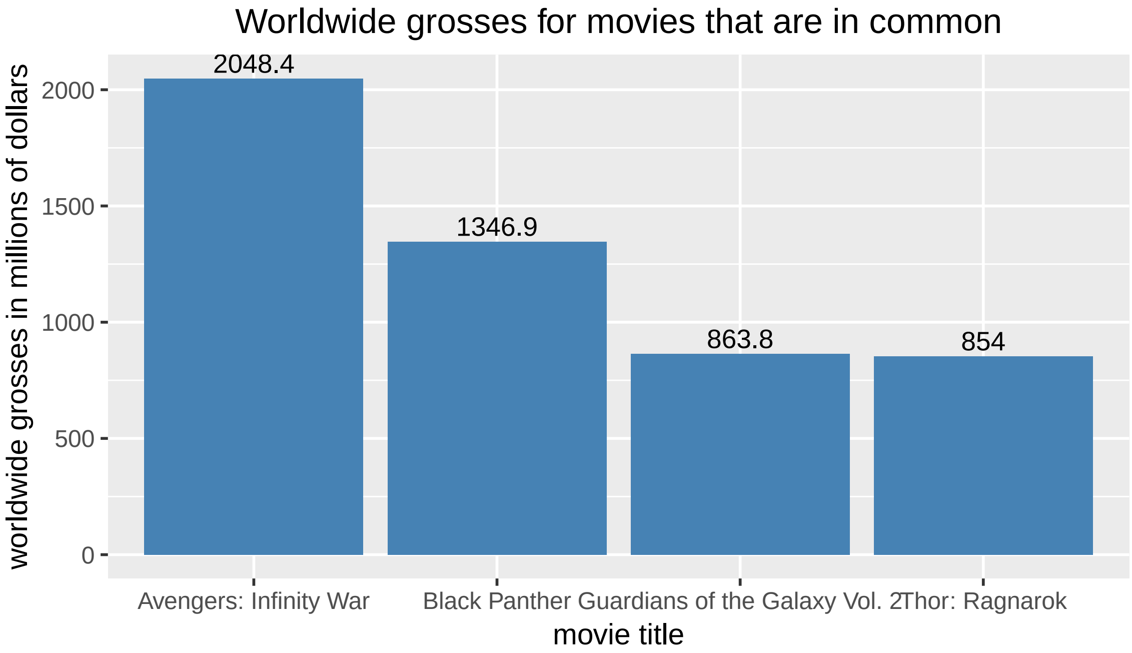
Netflix, IMDb, and Box-Office

Lastly, merge the all three datasets "netflix", "imdb", and "box_office" to get the common movies throughout the whole. Below is the result of merged datasets that gives 4 common movies. Interestingly, these four movies are all ranked in top 4 of the "netflix" datasets. Another interesting result is that fourth movie "Thor: Ragnarok" has ranked 43 higher than "Guardians of the Galaxy Vol. 2" in IMDb's rank, which is quite a huge gap. Out of these four movies, two are release in 2017, and other two released in 2018, which are all recent ones. Ratings for these movies are all above 7.0 and but below 9.0.

```
##                             Title Year IMDb_rating IMDb_rank
## 1         Avengers: Infinity War 2018         8.5         2
## 2                  Black Panther 2018         7.3        35
## 3 Guardians of the Galaxy Vol. 2 2017         7.7        42
## 4                 Thor: Ragnarok 2017         7.9        14
##   Box_office_rank.x Worldwide_gross.x Netflix_rating Netflix_rank
## 1                 5            2048.4            8.5            1
## 2                10            1346.9            7.3            2
## 3                67             863.8            7.7            5
## 4                70             854.0            7.9            3
##   Box_office_rank.y Worldwide_gross.y
## 1                 5            2048.4
## 2                10            1346.9
## 3                67             863.8
## 4                70             854.0
```

Bar Plots

According to the result of the common movies throughout all three websites, let's see the bar plot of worldwide grosses for each of the movies. It clearly shows "Avengers: Infinity War" is highest among all, and it has 2048.4 millions of dollars of worldwide grosses.



Most common non-stop words in Titles

Let's analyze the words in movies' titles. We can find out which words are more frequently used by taking out the stop words and then get the frequency of the words. First see which words in IMDb's dataset. When analyzing, the words in title have to be splitted and then create a new column that only contains each word. We can use "unnest_tokens", which is R function to split the words. Some of the titles may contain numbers, so take the numbers out using "-grep". Then anti-join with stop words will generate a column that only contains non-stop words of the titles. R contains stop words data as "stop_words" so I don't have to import stop words to analyze. Once anti-join with stop words, use "count()" function to get the frequency of the words.

Top 10 non-stop words of titles from IMDb's dataset

```
## Joining, by = "word"

## # A tibble: 10 x 2
##    word        n
##    <chr>   <int>
##  1 avengers     4
##  2 captain      4
##  3 dark         4
##  4 spider       4
##  5 america      3
##  6 iron         3
##  7 thor         3
##  8 ant          2
##  9 black        2
## 10 chapter      2
```

Top 10 non-stop words of titles from Box-Office's dataset

```
## Joining, by = "word"

## # A tibble: 10 x 2
##    word          n
##    <chr>     <int>
##  1 harry         8
##  2 potter        8
##  3 star          6
##  4 wars          6
##  5 spider        5
##  6 age           4
##  7 avengers      4
##  8 caribbean     4
##  9 pirates       4
## 10 dark          3
```

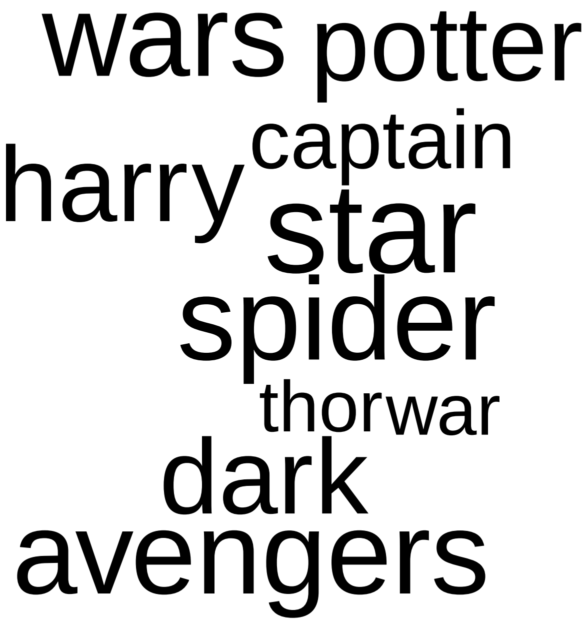
Top 10 non-stop words of titles from Netflix's dataset

```
## Joining, by = "word"

## # A tibble: 10 x 2
##    word         n
##    <chr>    <int>
##  1 matrix       3
##  2 black        2
##  3 jones        2
##  4 legend       2
##  5 star         2
##  6 summer       2
##  7 time         2
##  8 wars         2
##  9 world        2
## 10 american     1
```

Let's combine all of these sources of non-stop words to get the overall top 10 non-stop words. Then we can make a world cloud of the result using "worldcloud()" function from R.

```
## # A tibble: 10 x 2
##    word         n
##    <chr>    <int>
##  1 star        10
##  2 avengers     9
##  3 spider       9
##  4 wars         9
##  5 dark         8
##  6 harry        8
##  7 potter       8
##  8 captain      6
##  9 age          5
## 10 black        5
```



According to this result, word "star" placed in the first, next is "avengers", then "spider". Just by looking at these 10 words all seems describing heroic figures and something powerful. I can bring a hypothetical conclusion that people are desiring of superpower to overcome the problem that is hard to resolve. I can think of the reason that these types of movies are the most popular among all the movies is that in real world there is no such superpower, so that people want to satisfy themselves by watching heroic movies.

Conclusion

The movie websites, which are IMDb and Netflix, describing the popularities have different results. Their rank of movies are differ so that only contains 11 common movies among top 100's popular movies, which is just about 10 percent. Also by fitting various linear plots, the results show that there is no relationship between rating and rank of the movies. For relationship between worldwide gross and movies' rate and rating, there is no enough evidence to prove the relationship between them. There might be no enough data to get the result. It would be better if there has large dataset to compute. However, IMDb website only contains the top 100s popular movies. Therefore, in order to see the better relationship between grosses and rating or rank using these website, movies should be chosen more randomly without the preferences.