# Does book success guarantee its adapted movie success?

## Team 2 Members:

* Purnima Chandel
* Naim Panjwani
* Callan Yan
* Tigran Zohrabyan

## Abstract

In this project, we explore if success of a book guarantees success of the movie adapted on that book. To ensure we have a feasible dataset to work on, we have restricted our search to genre of science fiction.

We identified 237 science fiction films that were adapted from books from Wikipedia. These titles were queried on goodreads.com and the OMDb API to gather user ratings for the books and films, and the revenues for the movies.

We observed a strong, albeit highly variable, relationship between book and film ratings (p=6.9x10-6, r2=0.110). Furthermore, highly rated books tended to be great money-makers, but again with a lot of variability (p=0.033, r2=0.23). And, importantly, we did not observe a relationship between movie ratings and movie revenues (p=0.44, r2=0.03).

We thus conclude that there is a tendency of highly rated science fiction novels to have box-office success, but there are many other factors not explored here that make that success highly variable.

## Introduction

For decades, readers repeated the same phrase when Hollywood adapted a beloved novel for the screen: “The book is better than the movie.” But use of that phrase has gradually faded, replaced by enthusiastic shouts on social media when Hollywood grabs the rights to a classic work of science fiction or a modern twist on fantasy.

Books often lead to movie adaptations, and so much the better for sparking a fire toward both industries -- if a movie gets more people to read, great; if avid readers get to see their beloved characters on the big-screen, that's a boon too.

#### Project Objective:

Our question is whether the science fiction novels’ ratings correlate to ratings of films. Also, is there a correlation between science fiction ratings or film ratings with revenue obtained from a film?

To answer these questions, we pulled a list of science fiction books and their average reader ratings and mapped it with movies based on these books and their viewers/critics’ ratings and box-office revenue.

## Methods

### Data Sources Used:

* Websites Scrapped: Wikipedia, Goodreads
* API Used: OMDB API
* CSV: Kaggle

We scraped Wikipedia to get a list of science fiction films that have been adapted from books. We found 237 science fiction movie titles that have been adapted from books. These film titles were then queried on the OMDb API database to obtain IMDB/ rotten tomatoes ratings and box-office revenue.

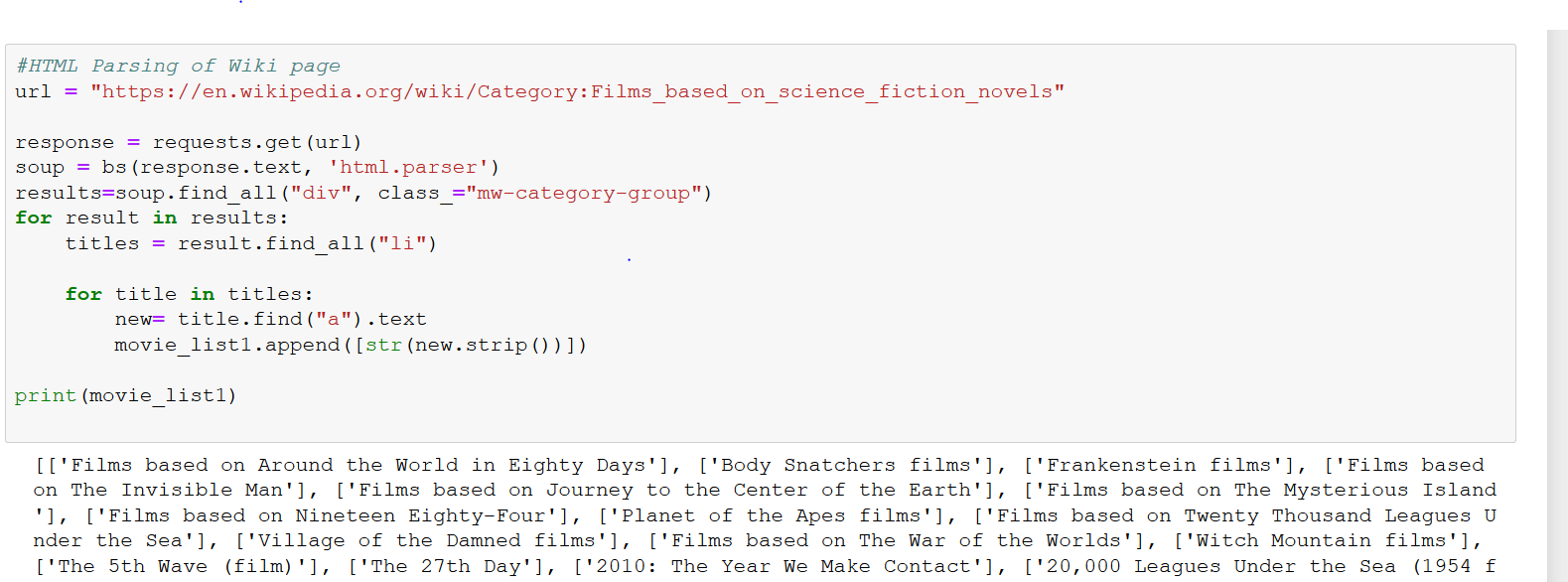
For book ratings, we scraped Goodreads for the list of book titles of interest and obtained the average user ratings for these books.

For the instances where one book was made into more than one version of a movie (e.g. one in the 80s and one more recently), we focused on the most recent film.

A more concise explanation with the links to all the python code is available on the GitHub README.md (<https://github.com/naim-panjwani/books_and_films>). Highlights of some of the code is given below.

#### Extract:

Step 1: Scrapping of Wikipedia to get the list of movies based on books:



Step 2: Querying the movie titles from Goodreads to get the corresponding books and their reader ratings:



Step 3: Querying the movies from OMDb to get the viewer ratings and box-office revenue:



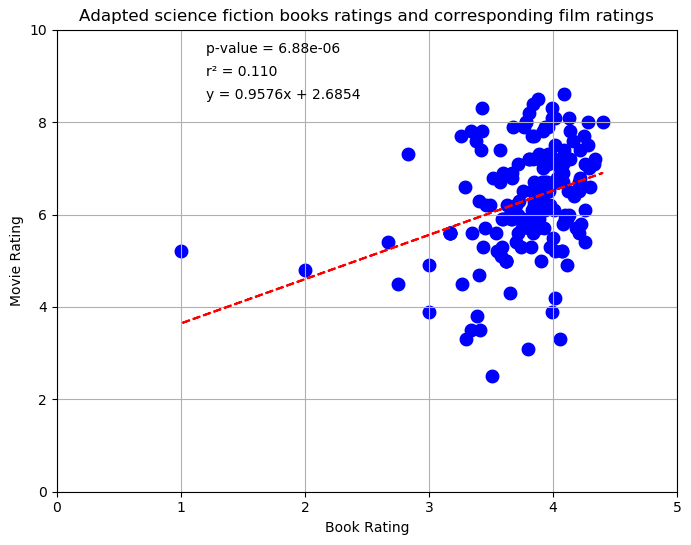
#### Transform

Step 1: Combine the data from different sources into one



Step 2: Run linear regression and plot various parameters



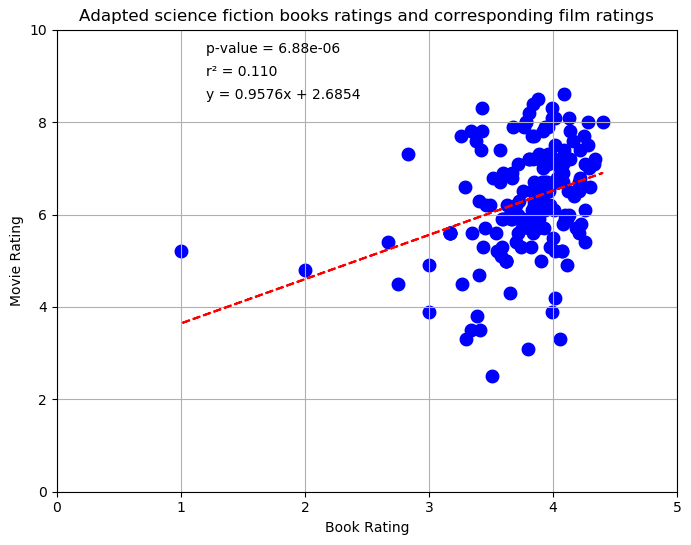


#### Load

Step 1: Load the final database into MongoDB

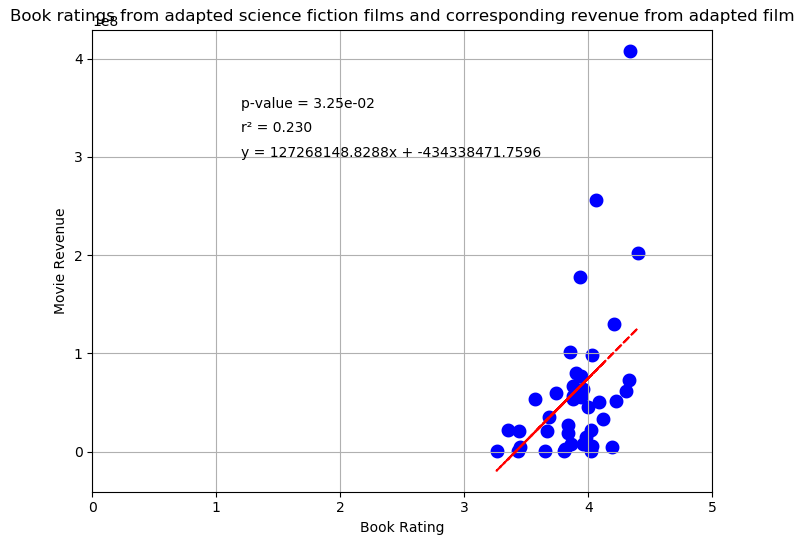


**Findings:**



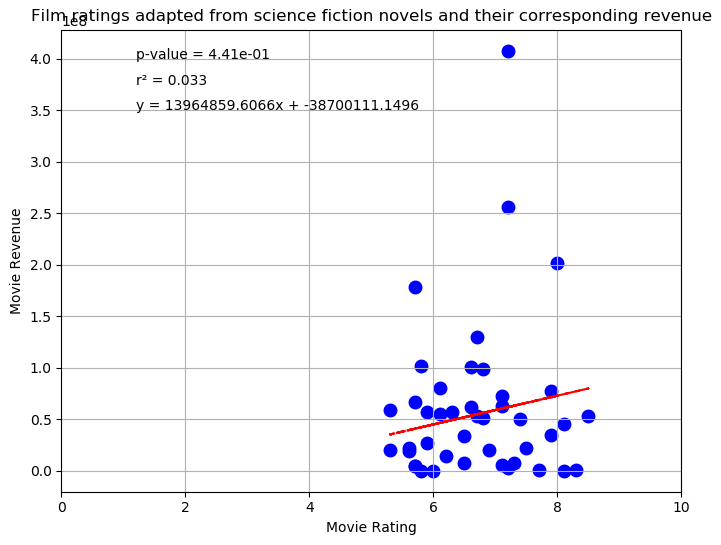
***Graph 1: Book ratings vs. movie ratings***

The p-value given is a two-sided Wald test (with t-distribution) for the hypothesis that the slope is zero. The p-value of book ratings vs movie ratings is quite low, which suggests that a relationship is present between book ratings and movie ratings. The r-squared value, however, is quite low suggesting a large variance between the two rating systems, or simply that the relationship is not exactly a linear one.



***Graph 2: Book ratings vs movie revenue***

The p-value for book ratings vs movie revenue is also low and suggests a relationship between book ratings and revenues from the adapted science fiction film. But, the low r-square suggests again suggests that we cannot fully explain the variance of the data points. In fact, the relationship appears to be exponential by eye!

******

***Graph 3: Movie ratings vs movie revenue***

The plot of movie ratings vs movie revenues shows no particular trend as indicated by the very large p-value and very small r-square value.

## Discussion and Conclusions

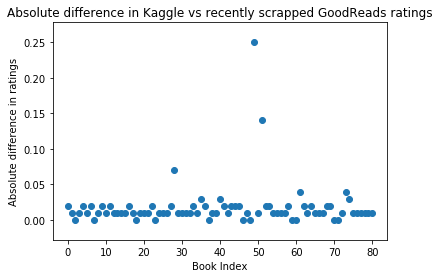
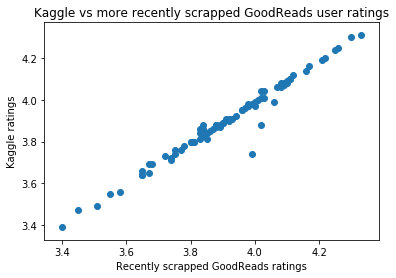
* We found a relationship between book ratings and movie ratings, which suggests that highly rated science fiction books do guarantee a successful movie adaptation to some extent. This relationship, however, has a huge variance and there are likely many other factors at play in the success of the movie such as production value, marketing, direction, etc.
* Interestingly, we found a relationship between book ratings and box-office success of the movie, but did not see any relationship between the ratings of the movies themselves with revenue. We caution, however, that this conclusion is derived from a limited number of data points

Some of the issues that we faced in our analyses were:

* Titles of the books and movies were different at times
* Box office data for all the queried movies were not found on OMDb
* Data cleaning took a lot of time as there were additional characters/spaces in the titles
* Scanning and identifying the right sources took bit of an effort. For example, we wanted to scrape Amazon for book reviews and add it to Goodreads reviews (doing a weighted average according to the number of user ratings) but Amazon blocked our scrapping queries after a few requests (and Chapters Indigo does not authorize scrapping of the user ratings from their website).

## Additional comment

We had a small interest in determining whether the user ratings in the Kaggle dataset released about a year ago (at https://www.kaggle.com/gnanesh/goodreads-book-reviews) were similar to those we obtained from web-scrapping GoodReads. You can see in the plots below that they are indeed similar for the titles that we were able to match:



***Graph 4: Comparison of our web-scrapped GoodReads ratings versus those obtained by Kaggle about a year ago.***

The plots show that the ratings have not changed much after a year for the book titles that we were able to match between the two datasets.