Project 3: Best Selling Books

Andrew Curtis, Carson Stoker

Introduction

The purpose of this analysis is to investigate the current trends and themes for best selling books. The New York Times publishes a popular set of best selling lists that ranks books within genres, subject matter and age groups. These lists are a common way that this information is shared with the public for the industry and in entertainment and news. The books are not only ranked by sales within the lists but their total time on the list in weeks is reported and is a popular metric to follow. A second dataset from Wikipedia is also used which lists the highest selling books of all time. This investigation looks to answer the questions: what factors make a book popular and have continued success? This will inform aspiring authors about what types of books to write if they want to be successful. This project utilizes many analysis techniques such as boxplots, scatterplots, pearson correlation coefficients, means, standard deviations, and t-tests. Several recommendations are elucidated for the purpose of identifying successful themes and avenues for developing a popular book.

Dataset

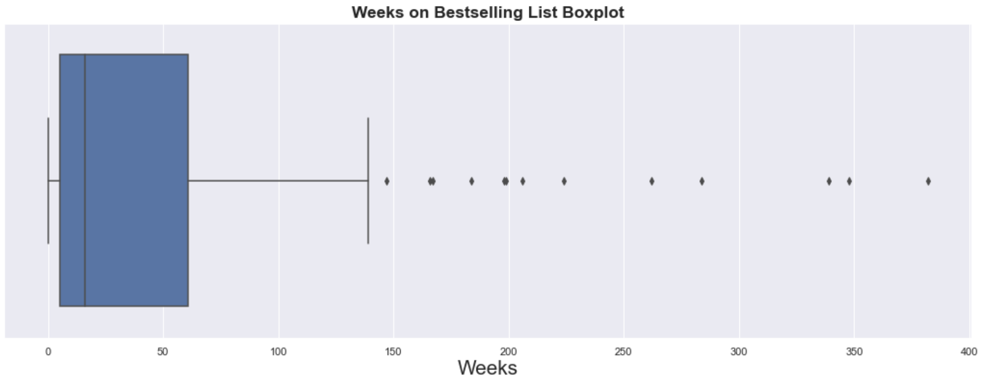
The New York Times Best Selling Lists are shown on their website. They have 11 lists in total ranging over a variety of subject matter, fiction or non-fiction and age groups. The data was web scraped from these lists. The fields contain information on: List and genre type, title, author, publisher, weeks on list, description and rank. The second dataset is from the Wikipedia page titled ‘List of best-selling books’. It provides the book title, author, publication date, approximate sales in millions of copies, and the genre. Having two datasets was useful as it allowed us to analyze the most recent, up to date trends, and historical sales.

Analysis Technique

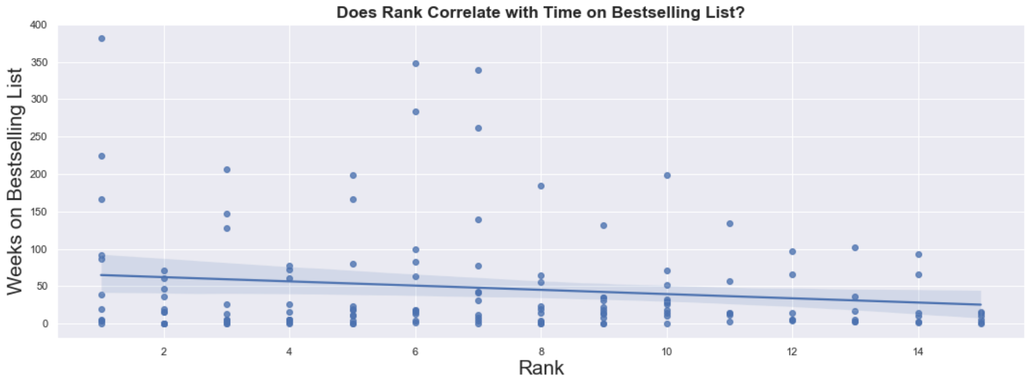
Several different methods for analysis were employed. Standard arithmetic metrics like mean and standard deviation are employed. This is always appropriate so as to understand the basic structure of the data. A box plot is given for weeks on bestselling lists. A scatterplot with a linear regression was computed for rank and weeks on bestselling list. A Pearson Correlation test was conducted for the same variables. A scatterplot and pearson coefficient were also used to relate time since publication and sales. This is appropriate because it gives us an idea of how two variables are correlated, which is important to understanding what factors make books sell well. Barplots for authors and publishers were found. Natural language processing was used on the descriptions to find the most common words and themes. A t-test was also employed to calculate if there is a difference between different types of books and how long they stay on the bestseller list. This is appropriate for this project because it gives us an idea of whether it matters what kinds of books an author writes.

Results

Investigation of the New York Times Best Selling Lists weeks on list field reveals how long a book stays near the top rankings and is an indicator of total sales. This variable was analyzed in several ways. A boxplot of this variable is shown below.

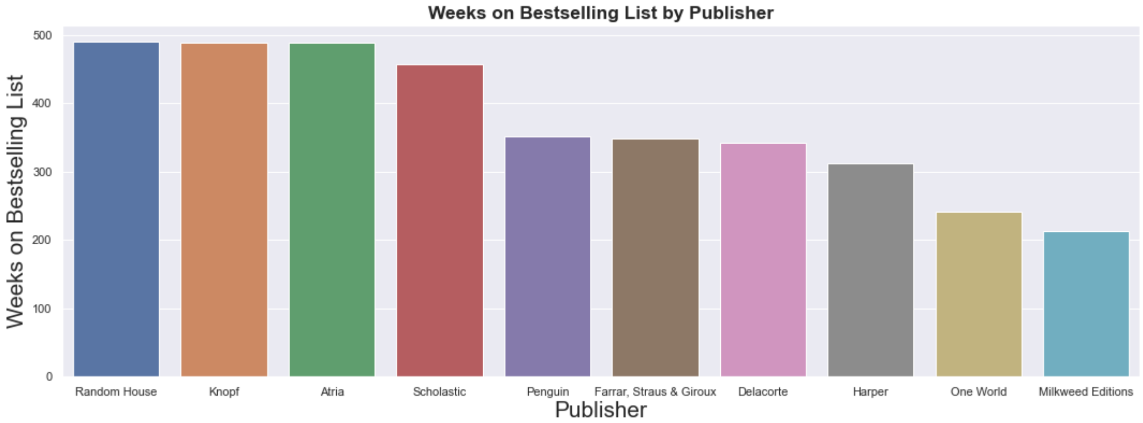


The average weeks on the list for a book was 47.6 with a standard deviation of 73.3. This is a large standard deviation compared to the mean for this type of data. However it is expected as there are some popular books that stay on the lists for large amounts of time. There is a rank assigned to each book in the list that is based on total sales. We chose to ask, do books with higher ranks tend to have been on the best selling lists for longer? This is answered in the figure below.



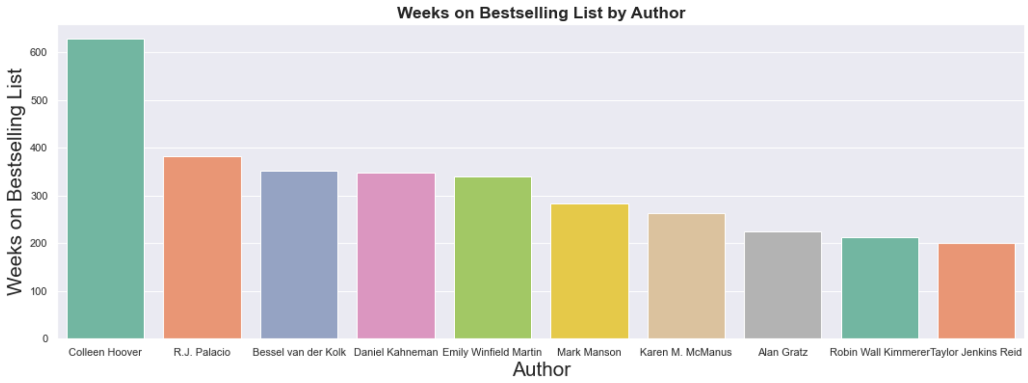
Indeed the rank of the book (1 being the top spot) is weakly correlated with being on the bestselling list for longer. Additionally, this conclusion was supported with a Pearson Correlation Coefficient test. The coefficient value was -0.157 with a p-value of 0.066. However, that fact may not be very useful since the correlation is so weak.

If you are a budding author, you may be looking to sign with a publishing company. We wanted to know which publishing company has the total most weeks on best selling lists. Below are the top 10 publishers.



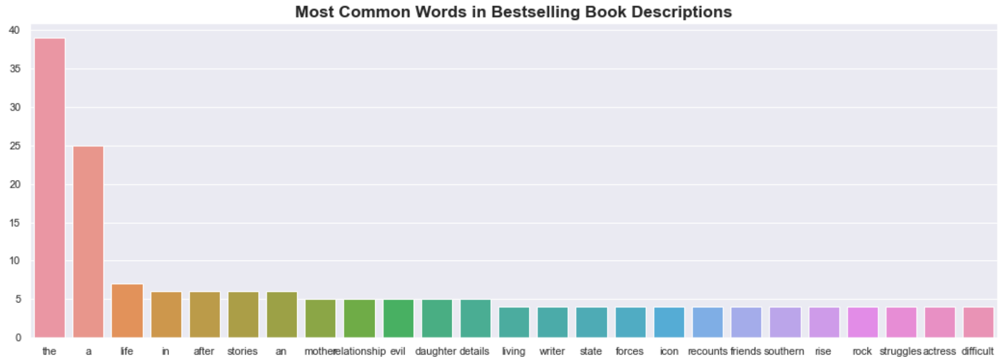
These are the publishers that get the most books on the bestseller list. It would be in the interest of new authors to publish through them. Random House is primarily a publisher of fiction books, while Knopf and Atria sell both fiction and non-fiction. Knowing these popular publishers may also indicate popular genres to write for.

Who are the most popular current authors and what genres are they working in?

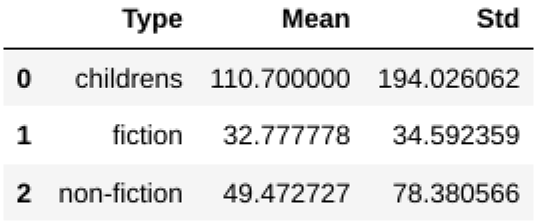


The top author with over 600 weeks on the lists is Colleen Hoover. She writes romance and young adult fiction. R.J. Palacio writes childrens and young adult fiction. Bessel van der Kolk writes on traumatic stress and treatments, fitting into the self-help genre. However, it should also be taken into account how large the market for a particular genre is. Some authors may have a large market share of a particular genre, and thus it would be difficult for new authors to find success in that genre.

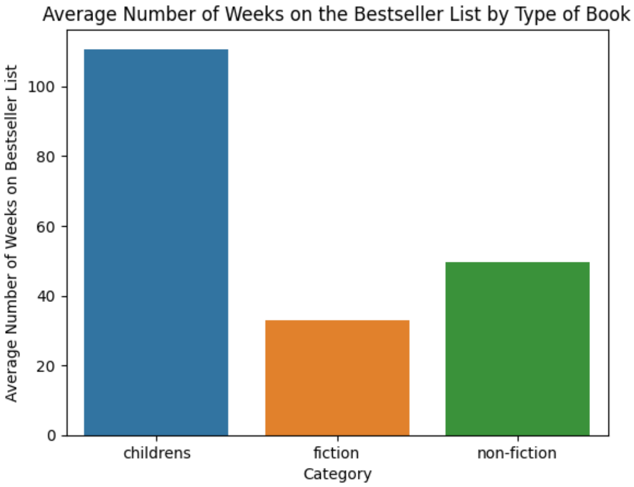
Descriptions of the books on the bestselling lists were joined and common words were searched for to investigate common themes and influences. This information can be helpful in determining trends for themes and subject matter within bestselling books. Below is a plot of the 25 most common words found after some natural language processing. Some interesting top results include: mother, relationship, evil, daughter, writer, southern, rock, struggles, actress. Many of the top words and books have themes centering around some hardship and also more of a focus on the female gender. This makes sense because in our experience, women read more than men do. So it makes sense that books marketed to women would sell better. It may be a good idea for new authors make women their target audience.



In the next part of the analysis, we grouped the lists by general category, following the template set by the New York Times. Shown below are the three categories and their average weeks on the bestselling list as well as standard deviation. Children's books tend to stay on the bestseller list for a long time. However, it should be noted how very high the standard deviation is. This means not all types of books stay on the list for long periods of time. We will explore this in more detail. Also shown below is a graph displaying the averages.



	Type	Mean	Std
0	childrens	110.700000	194.026062
1	fiction	32.777778	34.592359
2	non-fiction	49.472727	78.380566

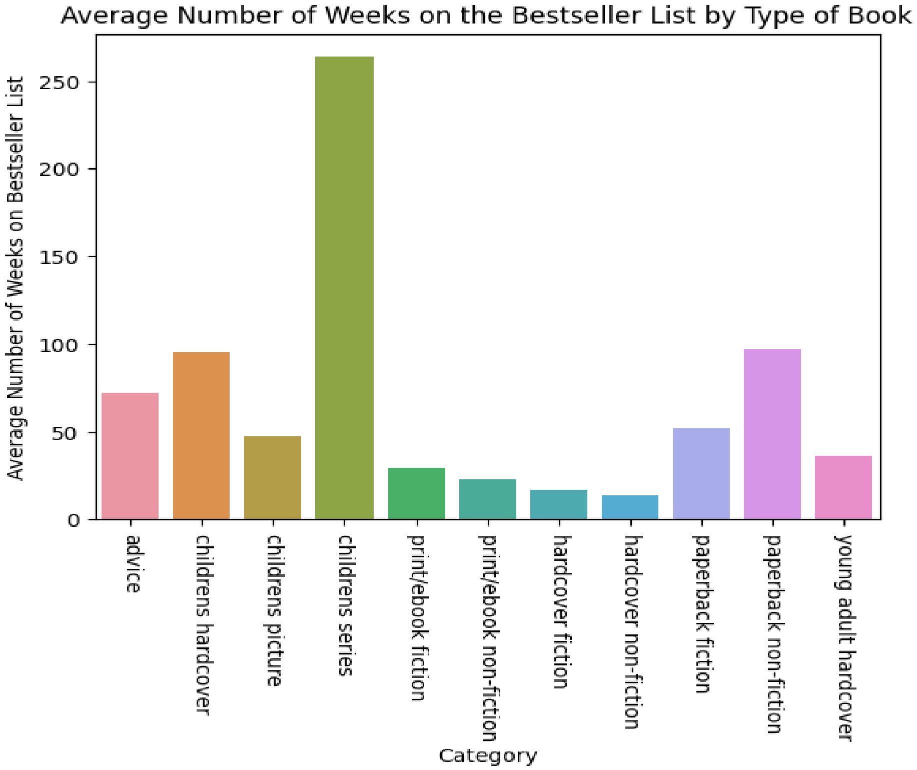


We should also mention that time on the bestselling list does not exactly equal the highest sales. Sales may vary greatly by category and since we only have relative sales, we can’t necessarily say that children’s books sell more copies than fiction and nonfiction books. All this tells us is that children's books on the list tend to stay there for long periods of time. This could be because there are less children's books being published and thus less competition.

Next, a t-test was performed between each of the three categories. It is no surprise, that the p-values for childrens/ fiction and childrens/nonfiction was low. We can already tell by the graph that there is a difference. The t-test for fiction/nonfiction found a p-value of 0.187. This means we can't be very confident at all that there is a difference in the number of weeks fiction and nonfiction books stay on the list. The results of these t-tests are shown below.

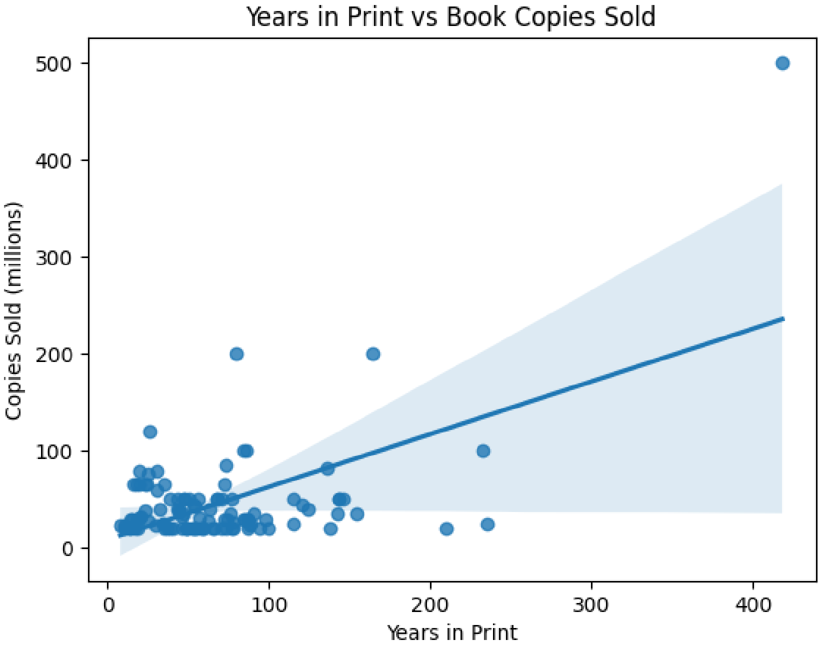
Types Used in Test	**T-statistic**	**p-value**
Fiction/Non-fiction	1.326	0.187
Childrens/Fiction	2.649	0.009
Childrens/Non-fiction	2.117	0.036

To further investigate how the type of book affects its time on the list, we separated the categories into the individual lists that the NY Times gives which also breaks down the categories into cover type. We can see that hardcovers stay on the list for shorter periods than paperbacks. This makes sense because hardcovers are more expensive. What's most informative about this graph is the outlier, children's series. In the graphs before, children's books stayed on the list much longer than fiction or nonfiction. This graph makes it clear that this is mostly due to the childrens series, which stay on the list about 2-5 times longer than the other types of books. This makes sense to us because children's series tend to produce a large number of short books that may stay relevant for a long period of time, leading to them staying on the list.

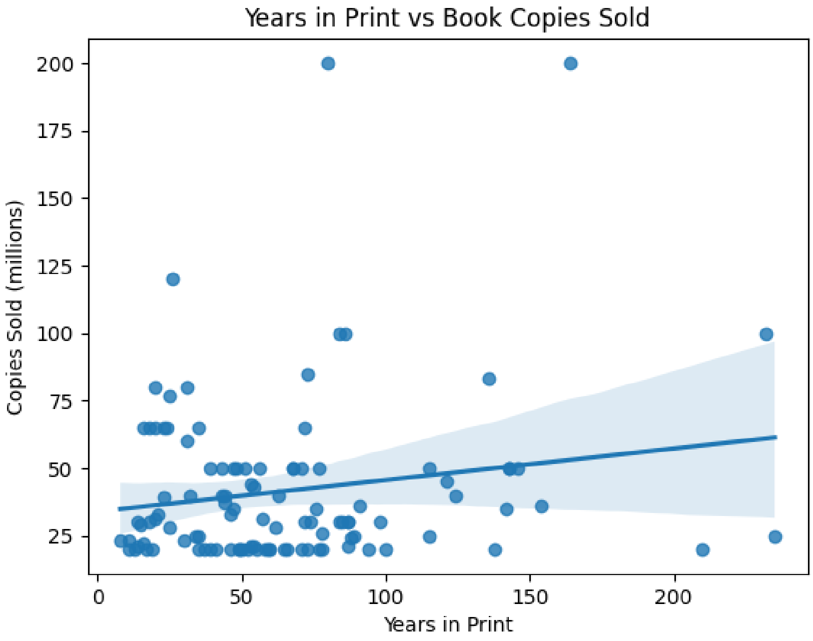


What we learn from this is that if authors want to stay on the bestseller list, they should write children's series. This may or may not actually correlate with absolute sales. The analysis previously showed there is a weak correlation between time on the list and relative sales. We acknowledge that this is a limit of this dataset, that it cannot tell us that information.

Finally, analysis of the wikipedia dataset was performed. A scatterplot with a pearson coefficient was made to graph the number of years a book has been in print vs how many millions of copies it has sold. The scatterplot is shown below. We found a coefficient (r) of 0.6116 and a p-vlaue of 9.029e-09. This tells us that there is a fairly good correlation and that we can be very confident this value is true. This tells us that the longer a book is around, the more copies it sells (at least for the most successful books. However, this does not necessarily indicate causation. This indicates if you want to be a successful author, writing something that will be relevant for a very long time tends to be associated with more sales.



Take note of the one large outlier in the top right. This is Don Quixote, which according to this dataset is the most well read book of all time. Since new authors are worried about making money now, and won't be around 400 years after their book is published, we decided to remove this datapoint and redo the calculations. The new graph is shown below. We found a coefficient of r = 0.1462 and a p-value of p = 0.2201. This tells us that when you discount Don Quixote, we can’t be confident of any correlation between how long a book has been in print and how many copies it has sold. Thus, this may be a factor that new authors do not need to worry about.



Technical

There were several outliers in the data set that were near 700 weeks on the list that were significantly skewing the data. These were removed for some of the analyses. They were mostly some children’s books that were common. It is reasonable to remove these not only to clean the analysis but also because they are not as indicative of current trends and are more holdover pop culture icons. There was a large amount of data cleaning and work involved with the web scraping. The natural language processing on the descriptions was mostly successful. Efforts were made to remove stop words and duplicates. As can be seen in the most common words diagrams there were a few common words that slipped through. Overall however, many interesting words were found that could indicate themes and subject matter.

The wikipedia dataset took less data preparation, but still required web scraping and putting all that information into a dataframe that was workable. The pearson coefficients were appropriate as it provided insight into how two variables are correlated, which is important to understanding what factors are associated with books that sell well. The scatterplots, means, standard deviations, and bar charts are also good ways to simply and easily convey insights about how data is organized. The t-tests were appropriate because they told if there was any difference in sales depending on the types of books that were written, so we can tell authors what kinds of books to write. We originally intended to more with the wikipedia dataset. However, it also had its limitations because the sale numbers listed are approximate. We wanted to do more analysis about genre and run t-tests but out of 100 or so books listed, there were 73 unique genres. This is because many genres were actually of list of genres and subgenres that the book fit into, making it too difficult to analyze this data without serious work to group the genres. In the future, we may have wanted to find another dataset that was better formatted and more informative.

Google slides link:
https://docs.google.com/presentation/d/1vAFATCgUUyXQl5dGUtW9p02b6Z6vdFT8ZqlBbMZsx60/edit?usp=sharing

Github repo link:
https://github.com/AndrewCurtis27/CS6830_proj3