Relationship Between the New York Times Bestseller and Amazon Best-selling Books

W200 - Summer, 2022

Theresa Azinge, Ana Zapata, Sunny Shin, Spencer Hodapp

07/31/2022

Introduction:

It's every author's dream to have their work published to the New York Times Bestseller list. Since 1931, it has been regarded as the gold standard in the publishing industry. To an author it cements credibility. To a publisher it signals good business sense. To the reader it signifies culture and a developed reading pallet. However, with the introduction of e-commerce and digital distribution of books, many strategic questions for the publishing industry have surfaced. Consumers are no longer relying on physical copies to read books. Therefore, publishing houses have witnessed significant growth in ebook adoption from the readers and an associated growth in ebook sales.

By comparing two sets of data, the New York Times Bestseller List 2010-2019 and Amazon Best-Selling Books 2009-2019, our team wanted to understand whether there is a clear correlation between the books that are listed on these lists as bestsellers.

We are hoping to answer the overarching research question, which is:

Are there common factors that make up a bestseller book? Based on the defined variables and our analysis can we predict what's the characteristics of the next bestselling book?

The variables we are drilling into are genre, year, number of reviews received, user rating, the duration of being a bestseller.

- Genre
    - What genre of books make it to the bestsellers list and how long do they stay on the list for? What is the genre with the most bestselling books?
    - How have the genres of books that make it to the list evolved over years?
    - Do the genres in the list change with seasons and how does it change?
- Dates
    - Is there a relationship between the release date and the duration on the bestseller's list?
    - What types of books last longer in time on the list?
- Number of reviews
    - Is there a relationship between the number of reviews received and the likelihood of a book being entitled as a bestseller?
- User Rating
    - Is there a relationship between the user rating for a book and the likelihood of the book being entitled as a bestseller?

Upon analyzing these variables, we are able to find whether there is any statistical relationship between the bestsellers on New York Times and most books sold on Amazon between 2010 through

2019. We anticipate that we would be in a better position to predict which books will be the next best seller per the New York Times' and Amazon standards.

Assumptions:

- **Ceteris paribus:** The data provided is a good indication of the total effect of the book market and is a holistic representation of consumer behavior.
- **Sales data will translate**: People buy books for a variety of reasons, but because we are looking at a decade-long scale, we assume that individual consumer preferences balance out over time.
- **Deliberately ignoring the effect of inflation:** Information around the prices of the book is only available between 2009-2013 for the New York Best Seller list. The team assumes that the 5-years' worth of data represents the market price of the books, and that the changes in price do not reflect the effect of inflation.
- **Amazon's sales data is indicative of books market standing:** Amazon started as a company dedicated to selling books, and as such they have profound industry knowledge on how to do so. Yet they utilize a dynamic pricing model to remain competitive across a variety of factors, meaning they are able to subsidize book prices in certain cases. Therefore, it is an important assumption to document going forward.

Dataset and Data Structure:

- New York Times Best Sellers 2010-2019[1]:
    - **Published_date:** This field shows the date when the book was published as a bestseller in the list.
    - **List_name:** This is the category of each book. Using this variable, we can analyze different types of categories on the Bestseller list by the New York Times.
    - **Rank:** This is the ranking on the list for that week when the book was listed as a Bestseller in the New York Times.
    - **Isbn10:** This is an International Standard Book Number (ISBN), which is a unique identifier for each book that consists of 10-digit numbers.
    - **Isbn13:** On January 1, 2007, the ISBN system switched to a 13-digit number format.
    - **Title:** This is the title of the book. Using this variable, we can analyze whether the same books appeared in Amazon's best-selling book list.
    - **Author:** This is the author of the original book.
    - **Description:** This field provides a short description of the book.
    - **Amazon_product_url:** The direct URL to Amazon's website.
    - **Price:** The price of the book.
    - **Weeks_on_list:** The number of weeks the book was on Bestseller list by the New York Times.

Out of these variables, we choose **List_name**, **Rank**, **Title**, **Price,** and **Weeks_on_list** as the most useful variables for analyzing the trend for bestselling books depending on the genre, time of the year, author, and prices.

[1] The dataset for the *New York Times Best Sellers* can be found here: https://www.kaggle.com/datasets/dhruvildave/new-york-times-best-sellers?select=bestsellers.csv

- Amazon Top 50 Best Selling Book 2009-2019[2]
    - **Name:** This is the title of the book.
    - **Author**: This is the author of the original book.
    - **User_rating:** The average user rating based on Amazon's 5-star rating system.
    - **Reviews:** This field indicates the total number of written reviews each book received. These reviews are written by the users.
    - **Price:** This is the average price of the book on Amazon, including the season discounts and promotional price-cuts as of October 13, 2020, inclusive
    - **Year:** The year the book ranked as one of the bestsellers on Amazon.
    - **Genre:** The books are divided in two general categories: Fiction vs. Non-fiction.

Out of these variables, **User_rating**, **Reviews**, **Author**, and **Price** were the most useful for analyzing the trend for bestselling books depending on the genre, time of the year, author, reviews and prices.

Additional Datasets Referred:

- Genre of each book[3]
    - This database was used to get a comprehensive knowledge of different and specific genres based on the fiction vs. nonfiction category.

List of fiction book genres	List of nonfiction book genres
1. Fantasy	1. Memoir
2. Adventure	2. Cooking
3. Romance	3. Art
4. Contemporary	4. Self-help / Personal
5. Dystopian	5. Development
6. Mystery	6. Motivational
7. Horror	7. Health
8. Thriller	8. History

[2] The dataset for *Amazon Top 50 Best Selling Book* can be found here: https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019?select=bestsellers+with+categories.csv

[3] The list for specific genre was consulted from the Self Publishing School website: https://self-publishingschool.com/book-genres/

9. Paranormal	9. Travel
10. Historical fiction	10. Guide / How-to
11. Science Fiction	11. Families & Relationships
12. Children's	12. Humor

The New York Times Best Sellers 2010-2019:

Focusing on the New York Times Bestseller list, we had taken the following steps to perform an analysis:

- Exploratory Data Analysis and clean data
- Analysis

The dataset was imported and decoded from the csv file.

First, as part of the exploratory data analysis and data cleaning exercise, a determination of irrelevant columns was determined, the following columns: *list_name_encoded, isbn13, isbn10, description, amazon_product_url* were removed as they would not provide any relevance to answering our research question. Next, we performed a statistical analysis of the data values of the columns to make sure outliers and irrelevant reports were not included in the analysis. It was determined that 69 author fields were left blank and were removed from the base data set.

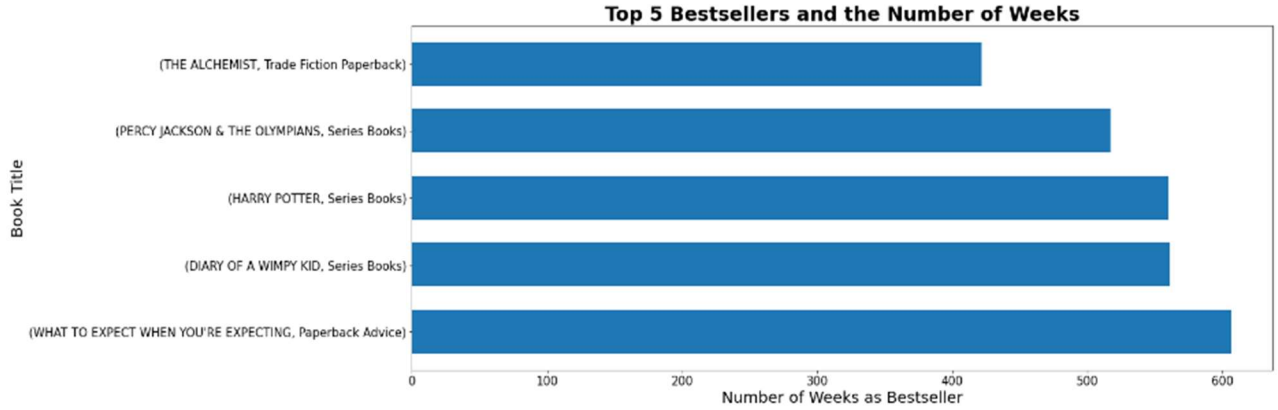
Next, we added several columns to the base data set, adding list_year and list_month to properly account for seasonality in our data. These were taken from the Published_date column. From this point, we wanted to check the data type for each field. It was determined that all accepted columns had acceptable data types for our purposes, and we could proceed with the statistical validation of our data.

From the statistical analysis, it was discovered that price information was not always included in books that experienced multiple weeks on the New York Times Best sellers list. This created a unique situation where we would have duplicate entries as a book continued to perform well but would under report price or would create erroneous entries. Normally, this would cause us to question the overall veracity of our data, but as this came from the New York Times itself, we determined to add a working hypothesis that price had fluctuated, and the NYT had elected to not track price changes. Therefore, it was decided to split the base data set into two bases, one that contained only one unique entry for each title, that preserved price, and another that recorded the dates, preserving the different rankins a title could achieve week to week.

Finally, after the initial data preprocessing was completed, we outlined the strategic analysis necessary to answer our research question and determine what common factors made a New York Times Bestseller. We started by trying to find commonalities and trends within groupings to give us a common starting point.

Top 5 #1 Bestsellers

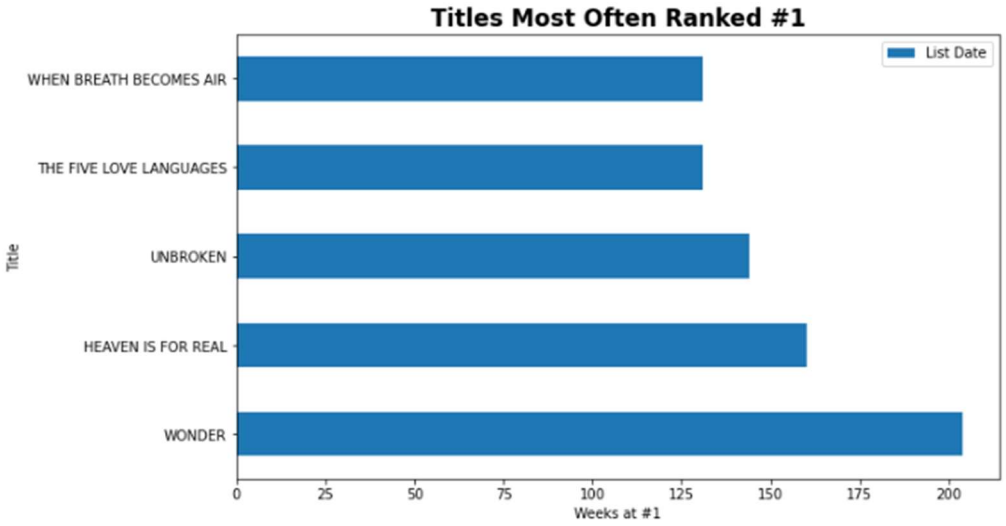
The most sensible metrics to start looking at the New York Times Bestseller data was to find out the top 5 bestselling books between 2010 and 2019. We took the number of weeks on the bestseller list as our variable and summed up the total number of weeks on the bestseller list per book. This method has two advantages: 1) we are able to find out that a particular book has been published to the New York Times Bestseller list once and stayed on the list for the longest period of time and 2) if a particular book appeared on the list *and* regained its fame by the public later in the year, we can gauge the book's total recognition over the last decade.



Based on our analysis, we found out that *What to Expect When You're Expecting* written by Heidi Mrkoff in 1984 was the top selling book of the decade. Based on weeks on the New York Times bestseller list, it is arguably the most influential book of the past twenty-five years. It also sold over 20 million copies worldwide and was made as a romantic comedy film in 2012.

Top 5 best-selling books that ranked #1 most often

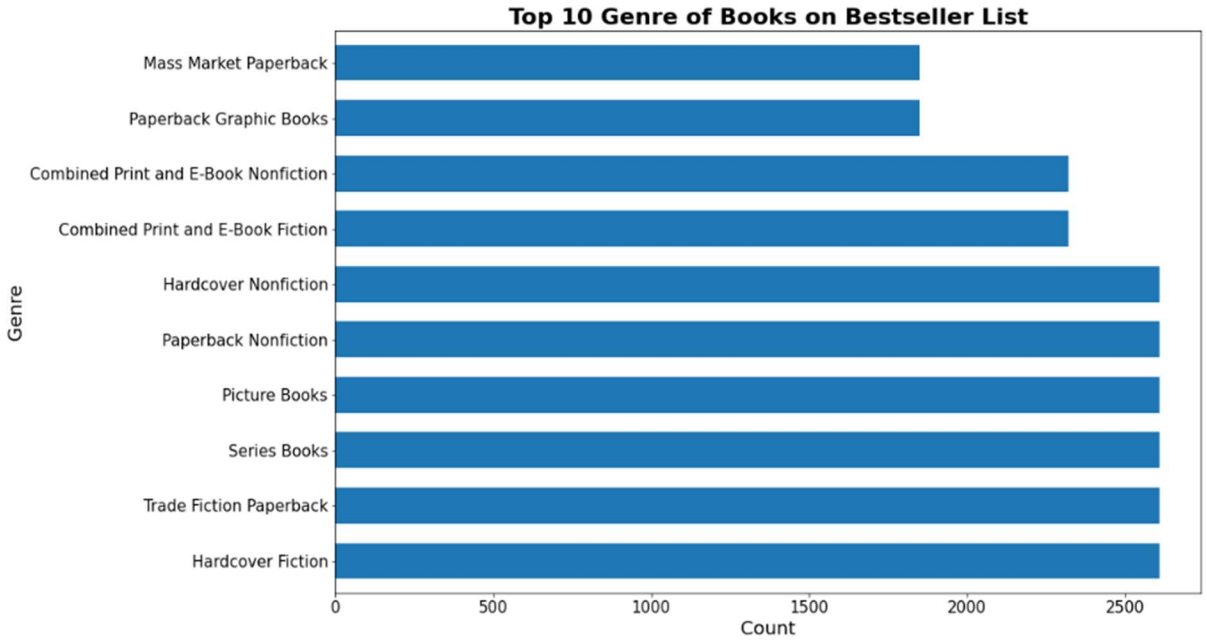
The next examination we performed was to look at books that dominated the number one rank in the New York Time Bestseller list for the last decade:



Wonder by R. J. Palacio and published in 2012 had the most successive weeks ranked number one at just over 200; it was categorized as a chapter book with a price of $15.99. The other four top books all came from different categories and had differing price points. This seems to indicate that the book themselves, not the attributions of the book, determine extreme popularity and purchasing power. In order for a title to remain at number one, it must not only continue to meet the minimum requirements of sales, but then beat out all other titles in the same category. Therefore, a book that remains at the number one spot must continually disperse out into the market and attract more and more customers, either through elite levels of marketing, high praise and or word of mouth marketing. Thus, it should be noted that a book that continually ranks number one, is an anomaly and does not accurately represent the average book.

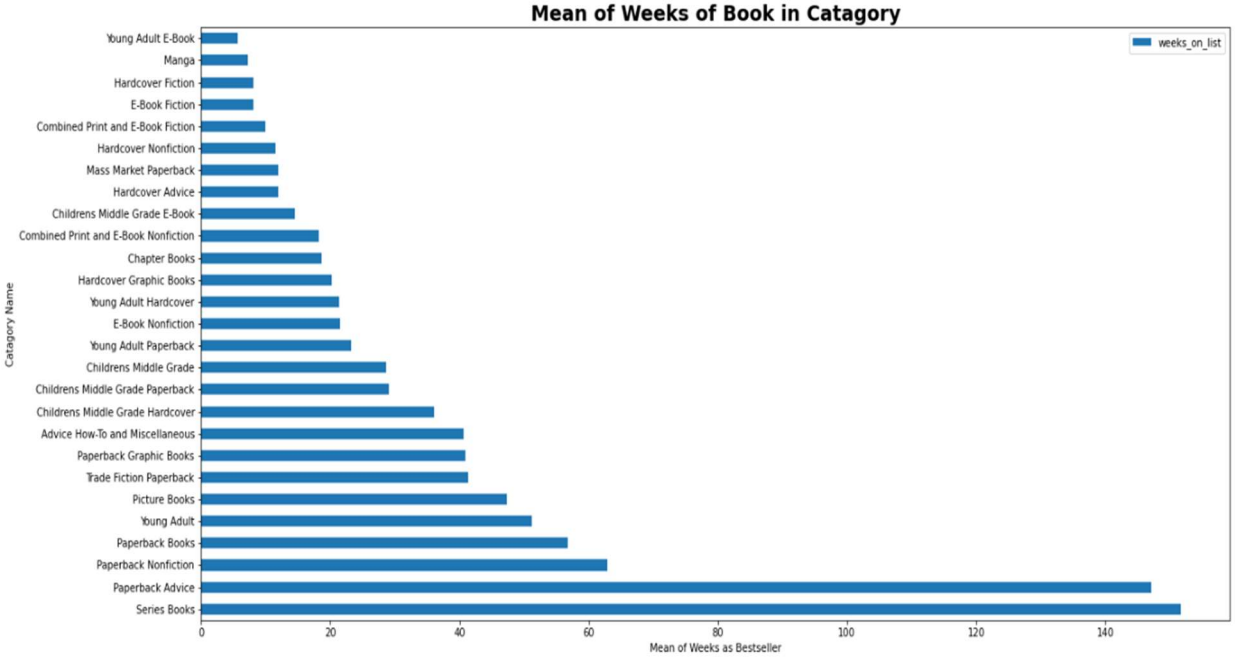
Top 10 bestsellers by genre

The next section of our analysis was to break down the top ten genres to see what book genre routinely performed well on the Bestsellers List. Unsurprisingly, series books took the top spot for a simple reason, a highly popular book with a sequel will probably help the sales of the sequel. In the Series sub-category, notable heavy weights like the *Harry-Potter* series, *Diary of a Wimpy Kid* and *The Hunger Games* were all sensations when they released. *The Hunger Games* in particular was notable for being ranked number one within the series category the longest.



Interestingly among the top 5 Genres, price stays relatively consistent with a mean of $17.99 and a standard deviation of $4.99. This seems to indicate that pricing as a strategy is well known among publishing houses as the only significant variation in price happens with a value addition to the product like a hardcover release or bundle. But taken as individuals it is germane to our question to understand that a tight pricing range significantly helps a genre's title perform.

We then took the mean of the number of weeks each genre of bestseller had to understand the average weeks each of the bestselling books remains on the New York Times Bestseller list.



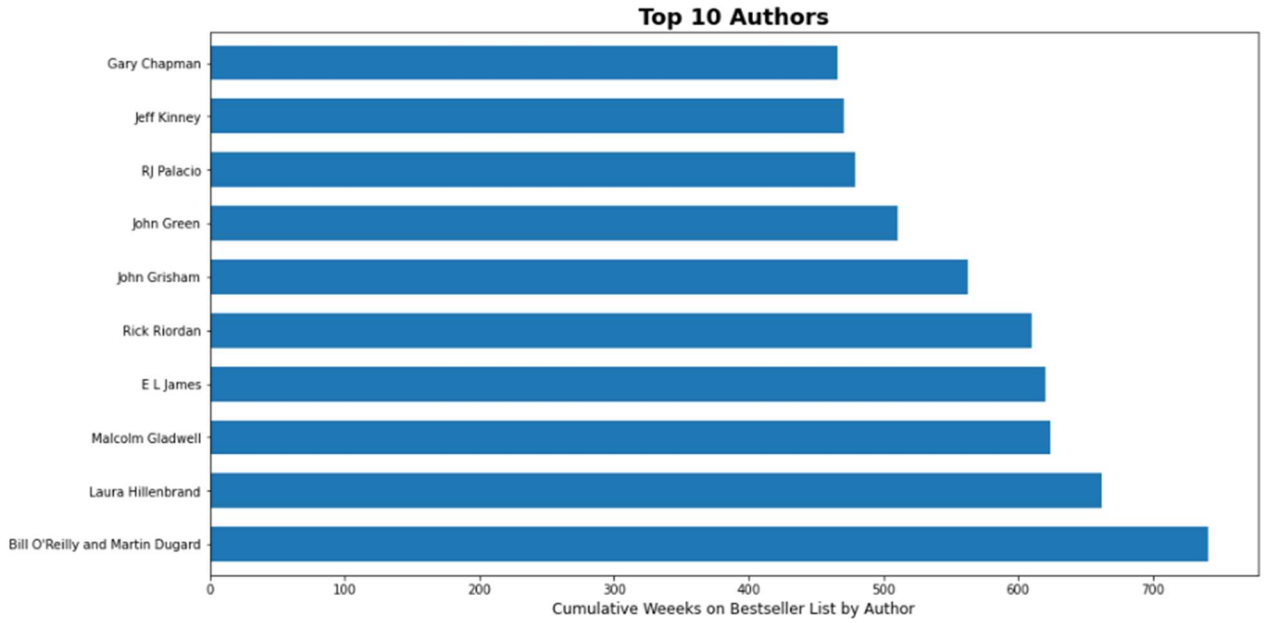
From this we were able to clearly see that the paperback advice genre, and series books, dominate the Bestsellers list and provide the best competitive advantage to an author seeking to get their book a long duration as a best seller. In fact, by looking at the duration of weeks as a bestseller we can clearly see that classes begin to form from 40-60 weeks then from 20-40 and a final group from 0-20. This would suggest strongly that there exists an upper bound on the longevity of a title based solely on its genre category. Further examination with different data will be required to determine the extent and firmness of this ceiling but it indicates that there is a general rule of thumb that exists per genre of title.

From this insight it was determined to run a Linear Regression for each Genre to better understand the role pricing played in overall book performance. For this we utilized the aforementioned unique data set so as to not introduce noise into our analysis. We ran a simple Linear Regression using the sci-kit learn package, with the following interesting findings found.

It was determined that rank was impacted by price, with every 50 cent drop in price a 0.43 increase in rank was determined. Additionally, there was an increase in duration for the weeks on the list with an increase of 1 week. Unfortunately, it should be noted that this analysis encompasses all genres, due to the asymmetry of the genre data it was not possible to determine genre specific regression results that meet our minimum R squared criteria of 0.80.

Top 5 best-selling authors

For this part we took a different approach and started grouping the books by authors. The main question we wanted to answer was whether there is any specific author who is writing a specific genre of a book that gets listed as the New York Times Bestseller. And if we can identify the author, can we also predict the number of weeks their book will be published as Bestseller?



Based on the analysis, we have learned that Bill O'Reilly and Martin Dugard had their books ranked #1 the most during the 2009-2019 decade. Specifically, the books they have written was the Killing Series (*Killing Lincoln, Killing Kennedy, Killing Jesus, Killing Patton, Killing Reagan, Killing The Rising Sun, Killing England,* and *Killing The SS)* which handles narratives of the true events surrounding the deaths and destruction of some of the most influential men and powerful nations in human history.

General Findings from the New York Times Bestseller List

In a general sense, if a book were to make it to the New York Times Best Sellers list, we would expect that the book should remain in the top five for its category for a duration of 24.7 weeks and have an average price of $19.85. Interestingly, the expected duration of a book changes drastically depending on the category of book that is written. For example, paperback advice books that cracked the list lasted on average of 124 weeks with an average price og $16.12 and the lowest category was Manga that lasted just 2.37 weeks at an average price of $10.59. It should be noted however that three categories lasted less than a week: Hardcover Business Books, Hardcover Political Books, and Paperback Business Books. After a careful continuation of the analysis, no particular culprits could be found for this trend in the data; therefore, it is our working assumption that these titles could not sustain the minimum threshold required to remain eligible for the list and were removed not to be displayed but due to a sudden contraction for demand.

Amazon Top 50 Best-Selling Books 2009-2019:

Focusing on the Amazon Best Selling Books list, we had taken the following steps to perform an analysis:

- Exploratory Data Analysis and clean data
- Analysis

The dataset was imported and decoded from the csv file.

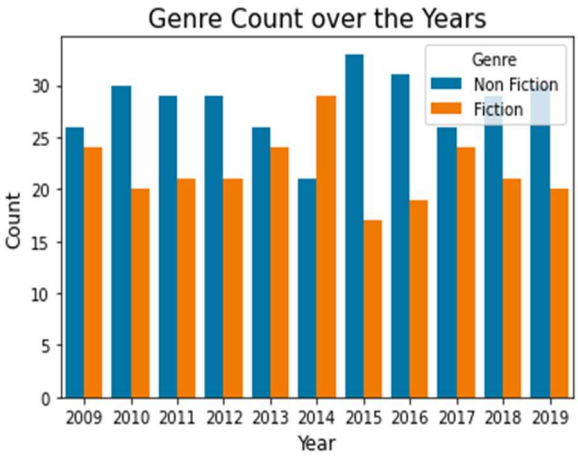
As part of the exploratory data analysis, the first step involved describing the data and cleaning it. This step revealed the following that all seven variables in the dataset (Name, Author, User Rating, Reviews, Price, Year and Genre) are useful for this analysis, so no columns were dropped. When we checked for anomalies and outliers the analysis revealed nine unique book titles with a price of $0. This anomaly could be explained by a discount, books that are listed at that price or an error in the data input. The dataset had no missing values or duplicate entries, with the variables properly titled and with the right data type. So, the dataset was used without additional cleaning for further analysis. There are 351 unique book titles written by 248 unique authors in the eleven-year period that the dataset covers.

From the statistical analysis, We noted two major differences with the NYT bestselling list. First, in the amazon dataset we discovered that the books are only categorized in two different genres: fiction and nonfiction. This created a unique situation where we would have to generalize categories and not specific genres for each book. Second, this list does show price information for each of the books presented. Meaning we could further elaborate our analysis using this variable and see for example the median price and how it fluctuates in the two genres.

Finally, after the initial data preprocessing was completed, we outlined the strategic analysis necessary to answer our research question and determine what common factors made an Amazon Bestseller. We started by trying to find commonalities and trends within each of the categories to give us a starting point.

Genres

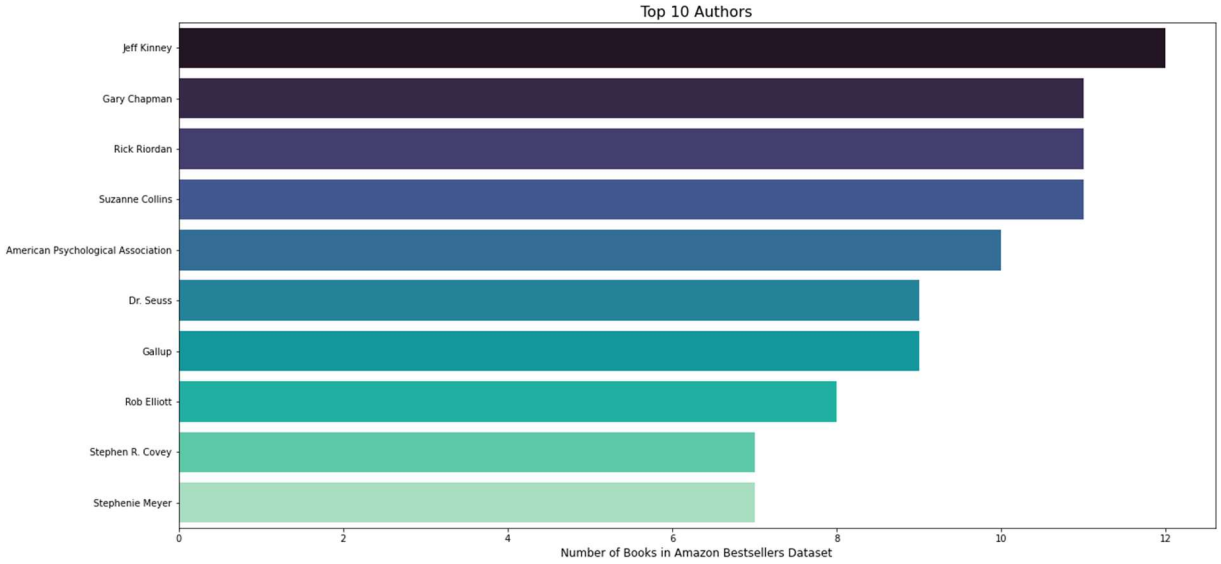
The first variable we explored was the genre. We took the genre variable and counted over the years variable. This method has the following advantages: We are able to find how the two genres relate over the years. Second, we are able to find out what is the genre with the bestsellers over the years.



Based on our analysis, we observe that a higher proportion of the books in the bestselling list each year is non-fiction with 2014 being the exception. Based on genre and the year's variable, In the eleven-year period that the dataset covers, more non-fiction books are present in the Amazon bestseller list than fiction books.

Authors

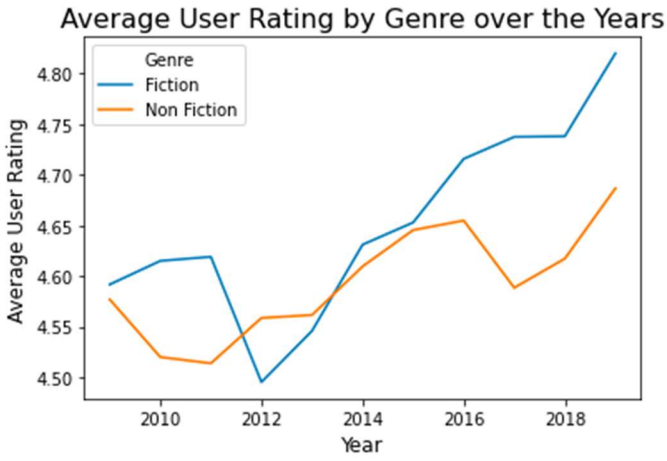
The next examination we performed was to look at the authors that dominated the list:



The top ten authors with the highest number of books in the Amazon bestsellers list are: 1) Jeff Kinney; 2) Gary Chapman, Rick Riordan, Suzanne Collins; 3) The American Psychological Association; 3) Dr. Seuss, 4) Gallup; 5) Rob Elliott; 6) Stephen R. Covey, 7) Stephanie Meyer. Further analysis indicates that each of these authors wrote books in only one genre class: Fiction. With fiction accounting for a slightly higher percentage of books written by the top authors. Authors like Dr. Seuss, Eric Carle, Gallup, Gary Chapman, American Psychological Association, and Stephen Covey, are top authors because their books made the bestselling list in multiple years. On the other hand, authors like Bill O'Reilly, Dav Pilkey, Jeff Kinney, Rick Riordan, Rob Elliott, Stephanie Meyer, are top authors because they wrote multiple bestselling books within the dataset timeframe. The authors who wrote multiple books that made the list were typically men and writing fiction series.

User Rating

The next section of our analysis was to look at books rating and how they performed in the last decade:

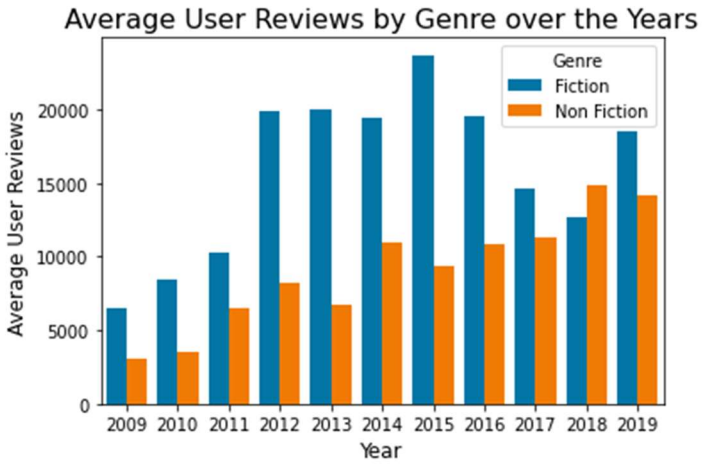


The User Ratings range from 3.3 to 4.9 with an average rating of 4.6 which is expected for books that are on the bestselling list. The books with User Ratings less than 4.0 are all fiction with *The Casual Vacancy* by J.K. Rowling having the lowest User Rating of 3.3.

The chart showing the trend of User Rating over the dataset period indicates that there is a slight positive increase in User Ratings over the years. Further analysis indicates that with the exception of 2012 and 2013, fiction books are typically better rated than non-fiction books. It is interesting to note that 2012 and 2013 also have the lowest average rating during the timeframe. The correlation coefficient between User Rating and number of reviews of approximately zero indicates that there is no linear relationship between both variables. The analysis also indicates a weak, negative relationship between ratings and price.

Reviews

The next section of our analysis was to break down the number of reviews for each bestseller:

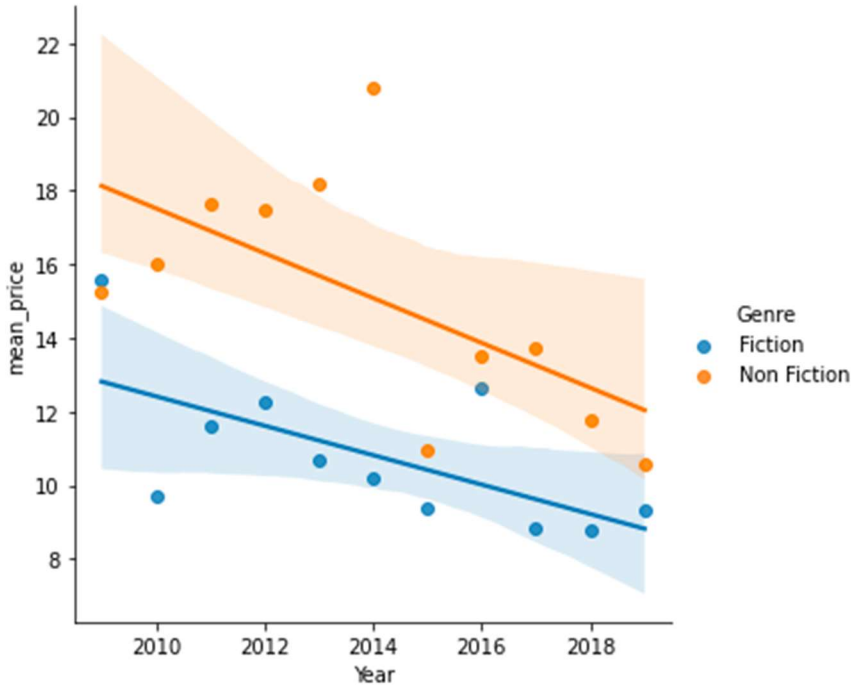


The number of reviews for each bestseller is within the range of 37 and 87,841. The books with lower number of reviews are mostly non-fiction, with *Divine Soul Mind Body Healing and Transmission System* by Zhi Gang Sha having the lowest number of reviews with 37 reviews.

Analysis of the data indicates an increasing trend in the number of average reviews over the years with fiction books getting significantly more reviews than non-fiction books except in 2018. The correlation analysis also indicates a weak, negative relationship between number of reviews and price.

Price

We finally look at the price ranges for both genres:



The range of books in the bestseller list is $0 to $105 with an average price of $13.10. Analysis of the data indicates that non-fiction books are more expensive than fiction books. However, the prices of fiction and nonfiction books decreased over the time frame. This could be explained by the introduction of the cheaper eBooks into the market.

General Findings from the Amazon Bestseller List

There are 96 books written by 90 authors that made the Amazon bestsellers list in multiple years. *The Publication Manual of the American Psychological Association* made the list for 10 consecutive years from 2009 making it the book on the bestsellers list the highest number of times. The top books over multiple years are also mostly Non-Fiction.

Books on the bestseller list for more than one year typically have higher User Ratings (4.63) than books that make it to the list only once (4.61). The average number of reviews for these multiple time bestsellers (15,957 reviews) is also higher than the average reviews for one-time bestseller books (11,953 reviews). The price of the books ($12.94) that top the chart multiple times is slightly lower than the price of the books that top the chart once ($13.10).

Conclusion/Next step:

After the previous analysis we noted that the duration of the data set encompassed a large period of time and should be noted that changes may have taken place to grading criteria or genre categorization. In looking for documentation tracking these changes it was discovered that the New York Times best sellers list has a minimum threshold of 1,000 to 5,000 books depending on the time period and that sales cannot come from one single entity to be included. This would indicate that previously it was possible to gamify a book into the bestseller list solely through purchasing power.

Interestingly it seems that no book was able to gamify the market and artificially keep itself on the best sellers list. Further digging into the New York Times criteria revealed that book sellers are often

surveyed by bookstores to cross-index findings with reports and surveys carried out by bookstores to hedge against a popular but bad book staying on the list. This shows a complex understanding of the role the New York Times Best Sellers list plays in the overall book market and is a further validation of the lists prestige and credibility when evaluating book sales.

Based on our findings between the two datasets, the team was able to obtain a few important data points which allowed us to formulate a recipe for the next best-selling book by both the New York Times and Amazon:

The New York Times Best Sellers 2010-2019:

- The book themselves, not the attributions of the book, determine extreme popularity and purchasing power. In order for a title to remain at number one, it must not only continue to meet the minimum requirements of sales, but then beat out all other titles in the same category.
- Series books took the top spot meaning, a highly popular book with a sequel will probably help the sales of the sequel.
- Among the top 5 Genres, price stays relatively consistent with a mean of $17.99 and a standard deviation of $4.99.
- The paperback advice genre, and series books, dominate the Bestsellers list and provide the best competitive advantage to an author seeking to get their book a long duration as a best seller.
- The rank was impacted by price, with every 50 cent drop in price a 0.43 increase in rank was determined.
- Bill O'Reilly and Martin Dugard had their books ranked #1 the most during the 2009-2019 decade.

Amazon Top 50 Best-Selling Books 2009-2019:

- Higher proportion of the books in the bestselling list is non-fiction with 2014 being the only exception.
- The top ten authors with the highest number of books in the Amazon bestsellers list are: 1) Jeff Kinney; 2) Gary Chapman, Rick Riordan, Suzanne Collins; 3) The American Psychological Association; 3) Dr. Seuss, 4) Gallup; 5) Rob Elliott; 6) Stephen R. Covey, 7) Stephanie Meyer. Further analysis indicates that each of these authors wrote books in only one genre class: Fiction.
- The authors who wrote multiple books that made the list were typically men and writing fiction series.
- The User Ratings range from 3.3 to 4.9 with an average rating of 4.6 which is expected for books that are on the bestselling list.
- With the exception of 2012 and 2013, fiction books are typically better rated than non-fiction books. The correlation coefficient between User Rating and number of reviews of approximately zero indicates that there is no linear relationship between both variables.
- An increasing trend in the number of average reviews over the years with fiction books getting significantly more reviews than non-fiction books except in 2018. The number of reviews for each bestseller is within the range of 37 and 87,841.
- The range of books in the bestseller list is $0 to $105 with an average price of $13.10.

Based on this data we can formulate that a bestseller for the New York Times and Amazon lists should:

- Be part of a **series**: Books that are a part of a fiction series dominate in terms of sales and longevity.
- **Rating**: We would expect a best seller to have a composite rating of 4.6.
- **Price**: Book will be priced between $13 - $17.99
- Bestseller **longevity**: Given this criterion we would expect the book to last on average 120+ weeks as a Bestseller.

Our research does not stop here; these datasets have a lot of interesting factors that can be extracted and further analyzed. One aspect of the New York Times Bestseller list that could take this research to the next level is the price data for all these books post-2014. Unfortunately, this specific dataset did not include prices for all the bestselling books published after 2014 as opposed to the Amazon bestselling list did include all the data. It made it difficult to perform a fair comparison between the two datasets. As a next step, the team can also try to bring in the specific genres of all the books on the Amazon bestselling list. Currently, the list only specifies whether the book is fiction vs. non-fiction. If we are able to obtain a more detailed level of genre for each book listed on Amazon bestselling books data, the prediction can be much more thorough and encompassing.

One of the hypotheses the team came up with but did not further test due to the time and resource constraints was regarding the population demographic that would primarily refer to these sources might be vastly different. We can argue that the population that decides to buy books from Amazon might be different from the group that pays attention to the New York Times Bestseller. We might not be comparing the same demographic. The New York Times looks at the qualitative factor whereas Amazon is geared towards the population who are about purchasing books at a more affordable price.

Specifically focusing on the Amazon bestselling books list, the team also believes that Amazon has an ability to sell ebooks, which may not be counted towards the total sales of the books. With the option of purchasing an ebook at much lower cost, the number of books purchased can be skewed on Amazon's list. The next step of the research can include another data point regarding the types of books on the bestseller lists and whether the New York Times Bestseller list also takes this into consideration.

References:

Max, Tucker. "How To Get On The NY Times & Every Other Bestseller Book List". *Scribe*. https://scribemedia.com/get-best-seller-list/

Pope, Bella Rose. "Book Genres: 24+ Genres for Writing (With Guides)". *Self Publishing School* https://self-publishingschool.com/book-genres/

Yucesoy, B., Wang, X., Huang, J. *et al.* "Success in books: a big data approach to bestsellers". *EPJ Data Sci.* 7, 7 (2018). https://doi.org/10.1140/epjds/s13688-018-0135-y