ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Exploring Factors Influencing the Amazon Best-Selling Books Selection Process from 2009 to 2019

Fatbardha Maloku* and Besnik Maloku

Master of Science in Business Analytics Student Candidates, Ageno School of Business, Golden Gate University, San Francisco, California 94105, USA

ABSTRACT

This research paper investigates the "Amazon Best-Selling Books" dataset, focusing on the methodology behind selecting top-selling books on the Amazon platform and employing analytical techniques to uncover key marketing factors. The study centers on evaluating consumer reviews and ratings collected from Amazon.com. It outlines the analysis methodology and discusses encountered challenges. Findings from the review analysis contribute to a discussion on influential factors in the Amazon book market, aiming to guide strategies for enhancing annual revenue from book sales at Amazon.

***Corresponding author**

Fatbardha Maloku, Master of Science in Business Analytics Student Candidates, Ageno School of Business, Golden Gate University, San Francisco, California 94105, USA.

Received: July 04, 2024; **Accepted:** July 08, 2024; **Published:** August 05, 2024

Introduction

Generally, it is assumed that the bestselling books depend on having great online ratings, but is it always true? Are there any additional elements or factors that have a direct or indirect impact on converting an online book into a best seller? The answers to these and other pertinent issues will be clear once we've completed the analysis. The goal of this project is to measure the progress and identify the top selling books in an online platform like Amazon, against historical data from 2009 until 2019. We will then predict the best-selling books based on rate scores and see how other external factors such as the customer reviews, comments, listed price, the social media influence, book cover and description play into this whole process. The Amazon.com is one of the largest online marketplaces, with millions of customers from all over the world. One of the features that this online marketplace offer is the ability to buy and rent books. A few attributes of this entire process are included within the "Amazon top best-selling books" dataset. This set contains variables such as the name of the book, the author of the book, the rating score, the number of written reviews, the price of the book, the year, and the genre it belongs. In this study, we'll investigate new approaches to increase purchasers' engagement with Amazon's book website, as well as giving suggestion and predictions on how Amazon can increase its yearly revenue from the book sales market.

Problem Statement

Business Problem Statement

Amazon makes roughly $386 billion in revenue each year, with book sales accounting for only about $5 billion.

Business Problem Background

Amazon makes just over $5 billion in revenue from their annual book sales. Surprisingly, this accounts for less than 10% of their entire sales revenue. This issue is caused by a variety of factors. It could be a low customer rating score on a book, negative reviews, the genre of books offered on Amazon, or the cheap/high price per book. In order to analyze this problem further we have selected a dataset of the top bestselling books in Amazon.com website. In this analysis, we will investigate innovative ways to improve buyers' involvement with Amazon e-book website, particularly among prospective customers as well as increase the yearly revenue of Amazon coming from the book sales market.

Desired Outcome Goal

- Regression Analysis Model on previous years to predict the top books rating.
- Identify the most memorable attributes or variables associated with the bookselling process
- Identify innovative ways to improve buyers' involvement with Amazon e-book website.
- Increase the yearly revenue of Amazon coming from the book sales market.

Model Selection

When searching through the data, I have looked at each variable with the following question in mind: Is there a recipe for creating bestsellers books in the Amazon.com website? In other words, what patterns I may uncover that might be influencing a book's or author's commercial success (as measured by sales volume)? Whatever pattern emerges from this investigation, it's likely that it's just a general trend for all novels that have also been bestsellers, rather than a distinguishing feature that made them bestsellers. It's also important to remember that past or historical data does not always predict what will happen in the future. These patterns may have evolved between 2009 and 2019, but there's no assurance they'll persist in the upcoming years. Taking all these elements into consideration the following models have been used and described in detail. This research paper will start by doing a **descriptive analysis** model on the historical dataset that we have gotten from Kaggle.com. We'll look at each attribute in depth and see what

some of the primary variables are that contribute to an Amazon book being a bestseller, as well as how the business can use this recipe to improve the number of book buyers on their platform. Following that analysis, we will use a **prescriptive analysis** model to identify the most memorable attributes or variables associated with the bookselling process. We will later continue by conducting a **predictive analysis** model on how the insights drawn from the previous analysis predict the sales volume and revenue of the bestselling books in the Amazon.com platform.

Solution Process

The analysis will start by doing a descriptive analysis of the first data set. Description of the "bestselling books" metadata is described as following. Here is the list of the columns that are available within this data set.

Descriptive Analysis

- Book Name – Name of the book
- Author Name – The author of the book
- User Rating Score – most relevant since it indicates how popular the book is
- Total # of Reviews – Total number of reviews on amazon books
- Book Price – The price of the book
- Year – The year on which the book is ranked as the bestseller
- Genre – The genre the book belongs to, whether fiction or non-fiction

The following is the definition of the rating scale:

- 4.0 - 5.0 = Outstanding
- 3.5 - 3.9 = Highly Recommended
- 3.0 - 3.4 = Recommended
- 2.0 - 2.9 = Disappointing
- 1.0 - 1.9 = Unpleasant

Distribution of the User Ratings Per Best-Selling Books

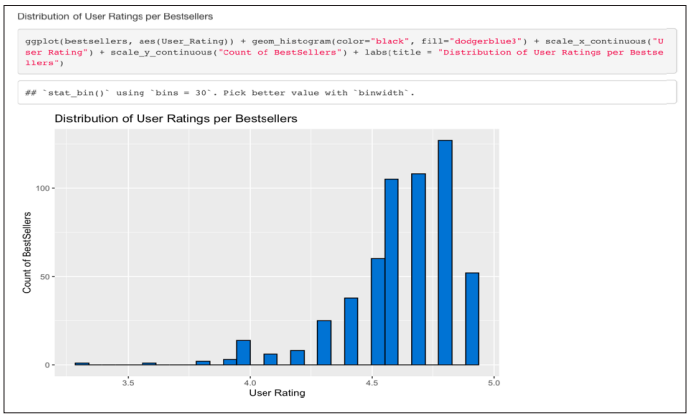

Figure 1: Distribution of User Ratings Per Best-Selling Books

There is some variance in User Rating, as shown in the graph above. A total of 180 books have received a 4.8 or 4.9 rating. Meanwhile, 9 bestsellers have received less than a 4.0 rating. The annual mean ratings, as can be seen, are fairly close, ranging between 4.5 and 4.8. As a result, we may claim that, on average, the quality of best-selling novels is guaranteed.

Distribution of the Reviews Per Bestselling Books

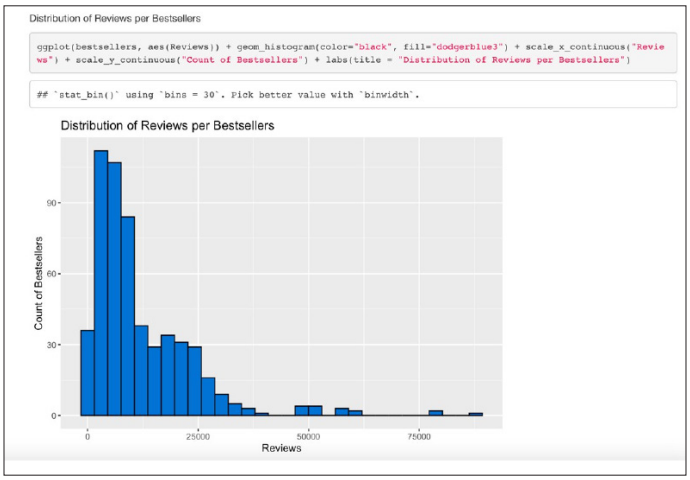

Figure 2: Distribution of Reviews Per Best-Selling Books

The majority of best-sellers appear to have fewer than 40,000 ratings in terms of reviews. In subsequent years, just a few of the books have remained among the best-selling titles.

Distribution of the *Price* Per Best-Selling Books

When it comes to the prices of book, the bulk of popular novels are around $20. From the chart below we can see that there are other prices listed with a few books being cheaper and a few a bit more expensive than $20, however we got an insight that the average of a best-selling books should be in average rank price for it to be purchased by a large number of customers.

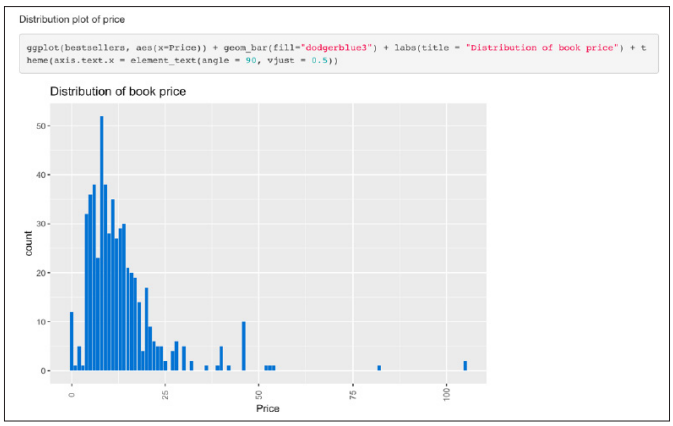

Figure 3: Distribution of Book Price

We investigated if the book's genre had any influence on the procedure after establishing the number of customer reviews, customer rating score, and price of the best-selling book. We've plotted the Price Average distribution over years for the "Fiction" and "Non-Fiction" genres below. The graph shows that the nonfiction category has been more crowded than fiction, despite the fact that the price point for nonfiction has been rapidly falling in recent years.

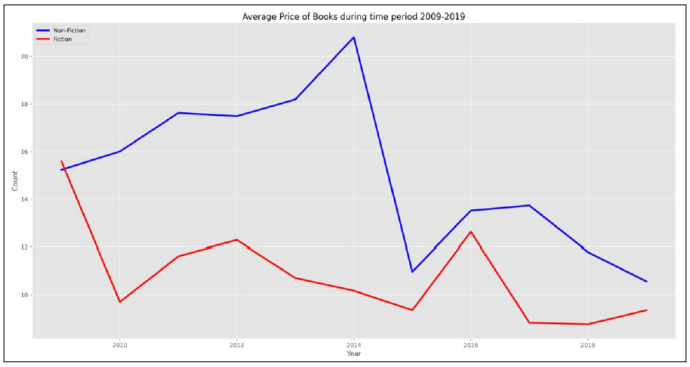

Figure 4: Average Price of Books during Period 2009-2019 by Genre

Below we have the distribution of User Rating Average over the years (2009-2019) for the "Fiction" and "Non-Fiction" genre. As shown, the average rating for the fiction genre has been quickly increasing in the recent years.

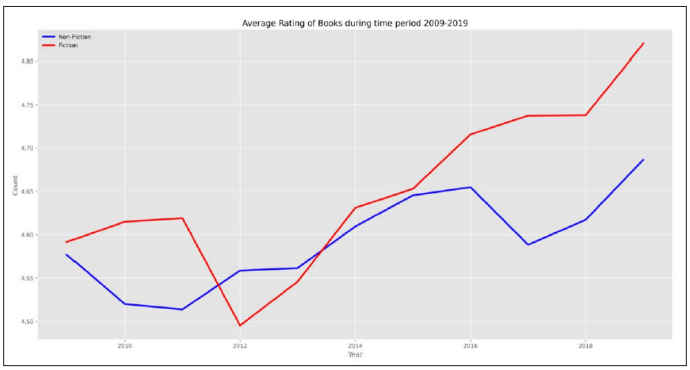

Figure 5: Average of Best-Selling Books during Time Period 2009-2019 Per Genre

The distribution of Reviews Average over the years (2009-2019) for the "Fiction" and "Non- Fiction" Genre is shown below. We can see that the reviews of each genre are not stable. The non-fiction genre per reviews is decreasing while the fiction genre is increasing.

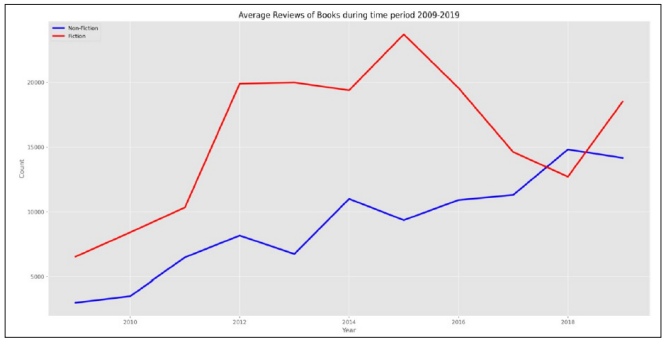

Figure 6: Average Reviews of Best-Selling Books over 2009-2019 Time Period Per Genre

The distribution of best-selling books by genre, year, and customer reviews comes next. As can be seen, the distribution of fiction and nonfiction is not constant, but rather varies from year to year. One thing to keep in mind is that genre fiction receives better reviews than nonfiction.

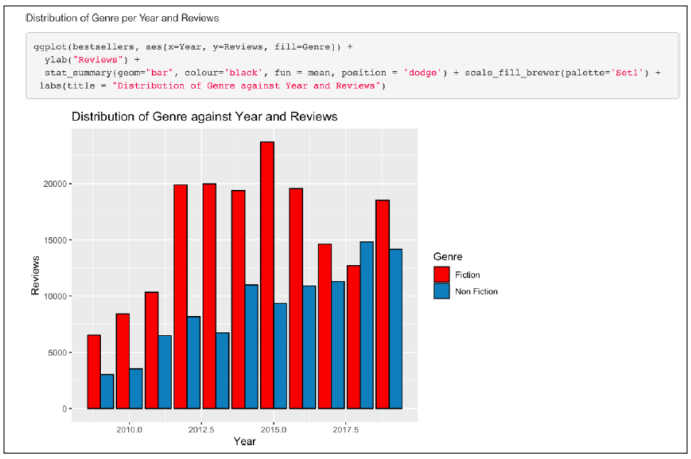

Figure 7: Distribution of Genre Per Year and Review Score

Below, we have created the Amazon's book market share in the US from 2009 until 2018.

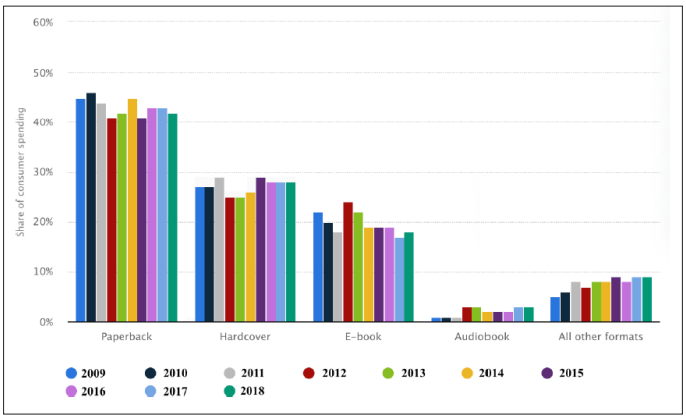

Figure 8: Amazon Book Market Share from 2009 until 2018

Predictive Analysis

The predictive analysis will be conducted using the below two criteria's:

- Correlation Analysis Between Variables
- Regression Analysis Model on Previous Years to Predict the Top Books Rating.

The correlation and regression model has been used to test the relationship between total number of books sold in Amazon and customer rating score, customer reviews, book price and genre. By exploring these potential explanatory variables, we will know what variables help make a book the bestseller and help drive traffic to the e-book Amazon page. The data we gathered is shown in the below table.

Reviews	Rating	Price	Year	Genre	Sales
4710	4.5	15.4	2009	0.1	487298000
5479	4.5	13.46	2010	0.1	497243000
8100	4.5	15.1	2011	0.1	464338000
13090	4.5	15.3	2012	0.1	487298000
13098	4.5	14.6	2013	0.1	410348000
15859	4.6	14.64	2014	0.1	450430000
14233	4.6	10.42	2015	0.1	437483000
14196	4.6	13.18	2016	0.1	458940000
12888	4.6	11.38	2017	0.1	438493000
13930	4.6	10.52	2018	0.1	494834000
15898	4.7	10.08	2019	0.1	486278000

Figure 9: Average Values for the Best-Selling Books in Amazon

We can run the correlation between these values using the data to evaluate how well or poorly they relate to the number of books purchased on Amazon. The explanatory variables will be perceived as statistically valid if they explain behavior at the 95% confidence level which we later look into once the regression model has been run.

Heatmap

Below, we have created a heatmap about the correlation between the number of best-selling books and other explanatory variables such as (Genre, Price, Rating Score, and number of reviews).

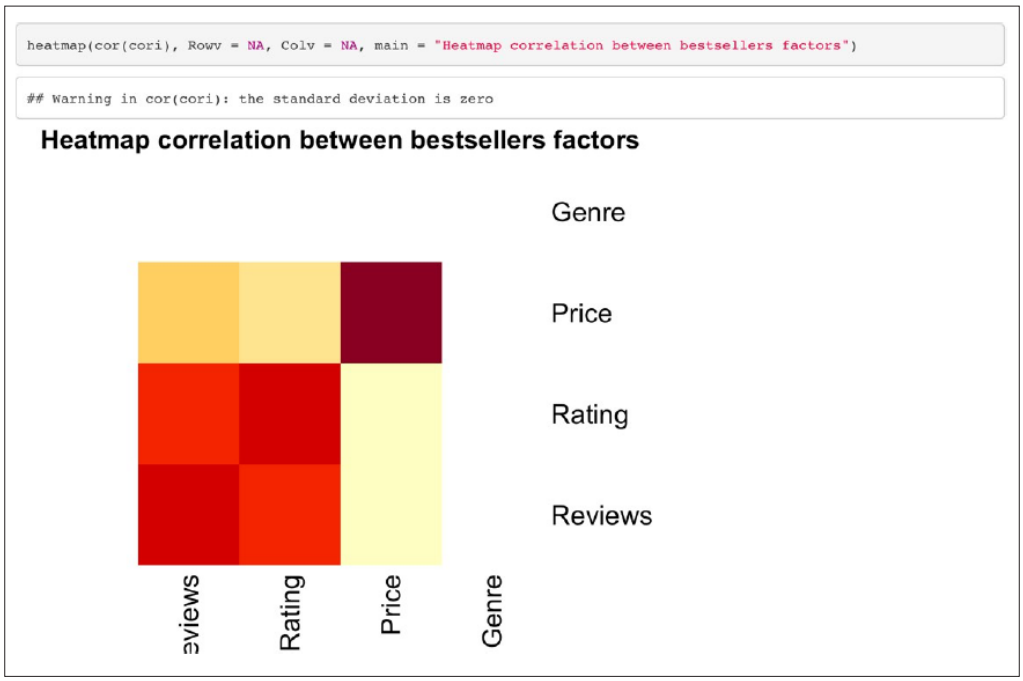

Figure 10: Heatmap between Explanatory Variables

As demonstrated in the heatmap above, the number of best-selling books is strongly connected with Reviews and Ratings, but less so with Price. We can also tell from the heatmap that genre is unrelated, as evidenced by the chart's white tint.

Best-Selling Book Factors Correlation

We generated a correlation chart with numbers to know the actual values of the correlation between the explanatory factors, as shown below. When the two variables (User Rating and Reviews) are analyzed independently, it becomes clear what makes a book stand out. The results also demonstrate that Reviews is a significantly associated variable with the number of books sold on Amazon. Just after the review with 0.72, we have the Rating score variable, which shows that when buyers buy books, they look at the book's rating score as well as the reviews. On the other side, the price is not substantially connected. With a total value of -0.49, the price connection with best-selling books remains negative. Finally, we can see that when purchasing books through the Amazon site, genre (fiction and non-fiction) is irrelevant for customers.

Figure 11: Correlation Numeric Values between Explanatory Variables

Regression

- Reason for Selecting the Regression Model
- This process will help us identify the most strongly correlated potential variables that will have influence over the website of the Amazon e-book visitors.
- This process will help concentrate our efforts on areas that will increase "Visitors" to the Amazon book selling website.
- Since e-book Amazon webpage has been decreased as the most relevant resource for sales and revenue for Amazon, it is important for us to focus on maximizing the daily visitors to Amazon.com and help drive the company revenue forward.
- With social media playing an important role in the Amazon e-book selling webpage, it is important for Amazon to listen to reviews and customer ratings and bring onboard books that will help not only the company review by driving customer traffic to the website but also make the page the most relevant one when it comes to purchasing books.

Model 1: Number of Books Sold ~ Reviews

```
sales = cori$Sales
reviews = cori$Reviews
model1 = lm(sales~reviews)
```

```
summary(model1)
```

```
## Warning in summary.lm(model1): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = sales ~ reviews)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.303e-07 -2.568e-08 -4.000e-11  3.036e-08  3.418e-07
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  4.873e+08  1.499e-07  3.252e+15   <2e-16 ***
## reviews     -2.725e-11  1.195e-11 -2.279e+00   0.0486 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.501e-07 on 9 degrees of freedom
## Multiple R-squared:  0.5247, Adjusted R-squared:  0.4719
## F-statistic: 9.937 on 1 and 9 DF,  p-value: 0.01169
```

Figure 12: Model 1 – Number of Books Sold ~ Reviews

Model 2: Number of Books Sold ~ Ratings

```
sales = cori$Sales
ratings = cori$Rating
model4 = lm(sales~ratings)
```

```
summary(model4)
```

```
## Warning in summary.lm(model4): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = sales ~ ratings)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.067e-07 -1.067e-07 -2.372e-08 -2.372e-08  4.863e-07
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  4.873e+08  3.832e-06  1.272e+14   <2e-16 ***
## ratings     -8.303e-07  8.396e-07 -9.890e-01    0.349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.79e-07 on 9 degrees of freedom
## Multiple R-squared:  0.4994, Adjusted R-squared:  0.4438
## F-statistic:  8.98 on 1 and 9 DF,  p-value: 0.01504
```

Figure 13: Model 2 – Number of Books Sold ~ Ratings

Model 3: Number of Books Sold ~ Reviews + Ratings

```
sales = cori$Sales
ratings = cori$Rating
reviews = cori$Reviews
combined_model = lm(sales~ratings+reviews)
```

```
summary(combined_model)
```

```
## Warning in summary.lm(combined_model): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = sales ~ ratings + reviews)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.405e-07 -2.664e-08  1.200e-09  2.887e-08  3.253e-07
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  4.873e+08  4.629e-06  1.053e+14   <2e-16 ***
## ratings      6.665e-07  1.047e-06  6.370e-01   0.5422
## reviews     -3.538e-11  1.778e-11 -1.990e+00   0.0818 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.553e-07 on 8 degrees of freedom
## Multiple R-squared:  0.4639, Adjusted R-squared:  0.3299
## F-statistic: 3.462 on 2 and 8 DF,  p-value: 0.04258
```

Figure 14: Model 3 – Number of Books Sold ~ Reviews + Ratings

Prescriptive Analysis

Prescriptive Analysis process will be conducted using the BI tools to simulate the desired business states by changing parameters to determine optimal business decisions. The idea behind this analysis is to identify the most memorable attributes or variables associated with the bookselling process. From the predictive analysis we got insight into the two explanatory variables (Reviews and Ratings) that have the most influence in the book selling process. The "What if" component is considered in the Prescriptive Analysis. In that case, our advice is to take a few steps that could shift the revenue of best-selling books to the Amazon.com marketplace.

1. Increase the number of good reviewers on book listing from the average of 8000 to 13000
2. Increase the number of customer rating score from the average of 4.5 to 4.7

We will undoubtedly increase our best-selling book revenues on Amazon.com if we investigate these two choices. Positive customer reviews of best-selling books increase brand awareness and commitment to Amazon's bookselling platform [1]. Customers are more likely to return as a result, resulting in increased sales and revenue for Amazon's bookselling platform [2]. If these two variables are taken into account, the number of books will increase as shown in the diagrams below

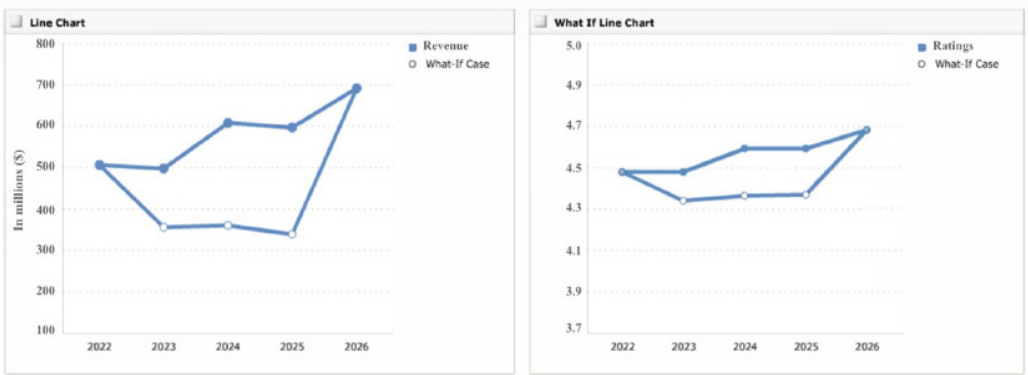

Figure 15: "What if" Visualization for the Amazon Revenue if Rating Score Increases Over Years

The next graph, we have an "What if" scenario of the revenue if the number of good reviews in the Amazon.com platform increases from 8000 to 13000.

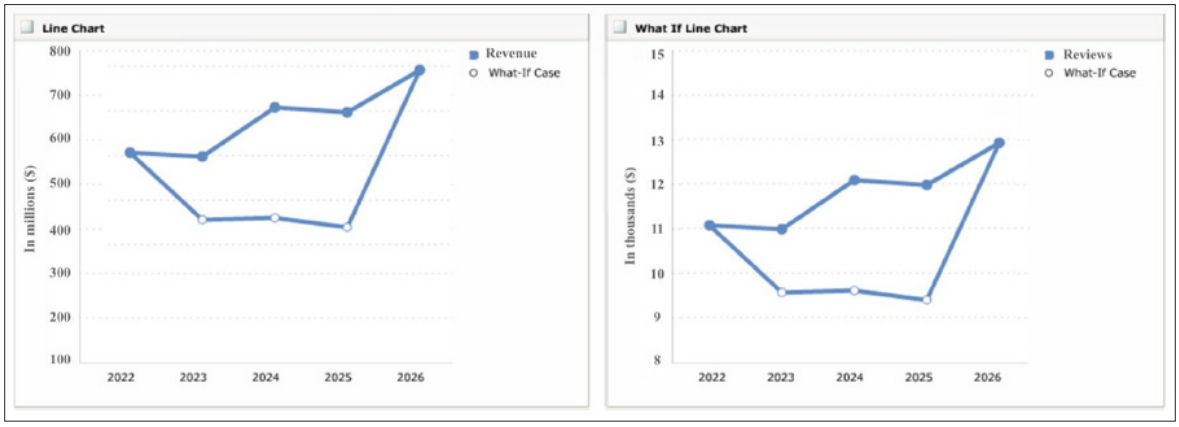

Figure 16: "What if" Visualization for the Amazon Revenue if Number of Reviews Increases over Years

As we can see from the visualizations above, changing the amount of customer rates and reviews has an impact on overall book sales, which is reflected in overall revenue of Amazon book selling list [3].

Literature Review

A literature review will be undertaken using a variety of peer-reviewed resources with findings or outcomes that are like mine. In this section, I will conduct research to identify other factors that influence the best-selling book category, either directly or indirectly. The main data set was found in Kaggle.com, however the whole research doesn't end there. I was able to use this dataset to continue my research and find additional internet sources that explored this topic further. By doing research I was able to find another data set that includes the "Amazon sales rank data for print and kindle books", as well as "Amazon Book Reviews" [4]. These two data sets contain information regarding the reviews that customers give to the book and the rank scores for the print books and kindle book. I will analyze these datasets and identify the main factors and then compare them to my main dataset and see how much of those factors influence the whole best-selling book process. There are several sources that I found which will help me in doing the analysis. All the resources are published, valid, and peer reviewed. Most of them were found in sites as ProQuest and GGU database library. Each of these sources contributes to my understanding of this topic by assisting me in identifying the aspects that have a significant impact on the book-selling process.

There have been multiple research articles done on the best-selling books topics. According to the "Price and non-price competition in the online book industry" article we understand that price doesn't have an impact on the book being a best-seller. By doing an analysis on the historical data of Goodreads the authors found that price does not influence the online booking industry [5]. Similarly, in the "Why Amazon uses both the New York Times Best Seller list and Customer Reviews" article we see that nowadays, majority of the products are reviewed online by the customers who use social media. The paper did focus on the relationship between the product sales which in this case are Amazon books and social media reviews from the customers. The research article found that social media reviews do influence sales volumes [6].

Other research articles such as the "Book reading behavior on Goodreads can predict the amazon best sellers" show that best-selling book do not just depend on the number of reviews and the number of rating scores. The team did an analysis on the Goodreads and did establish a connection between it and Amazon best-selling books. From the analysis, they concluded that the success of a book becoming a best seller depends on many factors. Their method was conducted based on the users posts and the genre of the book. The results of their analysis showed an improvement of 16.4% of the genre and users posts for the best-selling books over the ratings and reviews score [7].

We can see from the above study that the number of reviews and the rating score have the most impact on a book's ability to become a best-seller. Even if the plan may differ from one firm to the next, these two factors continue to be the most important in assisting books in being successful and increasing their market sales.

Model Results

Described below we have the results of regression models run for the dataset. In each one these three models we are going to conclude the model results and these results influence the overall book sales of Amazon platform.

Model 1: Number of Book Sold ~ User_Rating

After we run the regression model of user rating score against the total number of the book sold at Amazon.com, we got a p-value of 0.01 and an R-squared value of 0.44. From this model we understand that the p-value is highly significant over the other explanatory variables in the dataset.

Model 2: Number of Book Sold ~ Reviews

We then run the regression model of reviews against the total number of the book sold at Amazon.com, we got a p-value of 0.01 and an R-squared value of 0.47. From this model we understand that the p-value is highly significant over the other explanatory variables in the dataset.

Model 3: Visitor ~ Reviews + User_Rating

Lastly, we run the regression model of reviews and user_ratings against the total number of the book sold at Amazon.com, we got a p-value of 0.04 and an R-squared value of 0.32. From this model we understand that the p-value is highly significant over the other explanatory variables in the dataset. We also understand that when the explanatory variable reviews and user rating gets combined, we have a higher significant value that fits very well our model and gets us closer to the best possible outcome result.

Results Interpretation

In this analysis, we have conducted a correlation analysis of Amazon book sales over the period 2009 until 2019 against the average customer ratings, the total number of reviews per book and other explanatory variables. The correlation value between the book sold and customer rating score returned correlation coefficient of 0.72. In the similar instance the number of reviews returned a correlation coefficient of approximately 1. Based on these values, we identify the top two explanatory variables (Reviews and Ratings), that help a book become a best-selling one. At a 95% confidence level, our p-value should be less than 5% in order to invalidate the null hypothesis. Examining the p-values from the model results section, we make the following interpretations:

Model 1: (Number of Book Sold ~ User_Rating)

- At a p-value of 1%, we satisfy the 5% maximum limit. We can confidently state that user ratings are an important factor of best-selling books.

Model 2: Number of Book Sold ~ Reviews

- At a p-value of 1%, we satisfy the 5% maximum limit. We can confidently state that number of reviews are an important factor of best-selling books.

Model 2: Number of Book Sold ~ Reviews

- At a p-value of 4%, we satisfy the 5% maximum limit. We can confidently state that number of reviews as well as the number of ratings combined are an important factor of best-selling books. The value still fits within our limit.

According to the findings, a marketable book should include the following characteristics:

- Have been in the top 50 for more than a year
- Have a minimum user rating of 4.5
- Expect to pay roughly $20 for it.

Conclusion

In conclusion, we examined the dataset we got from kaggle.com. and utilized strategy tools to characterize the current state of the market and the elements that could influence the top selection of the books in the future. From the analysis we can draw a few definitive conclusions about what creates a bestseller. From 2009 to 2019, there was no single route for an author to be on the bestselling list. Others had books that sold well over several years, while others had series that became bestsellers over several years. Meanwhile, the rating and review profiles of fiction and nonfiction blockbusters during that decade were quite different. Nonfiction blockbusters won by a landslide in terms of number of reviews, while usual fiction bestsellers received higher ratings and won by a landslide in terms of number of reviews. Perhaps the expected rating and review profile for a bestseller differs between fiction and nonfiction works. The ratings and reviews in these datasets show customer behavior and feedback on our products. Based on the analysis, price, genre, and author popularity have a minor impact on customer behavior and might be regarded factors to encourage sales. According to the data presented above, Amazon will continue to be one of the most popular eCommerce websites for book purchases. By not only retailing books but also generating and selling their own, Amazon's Kindle Direct Publishing business puts them ahead of rival bookselling websites. According to the data above, Amazon's book sales revenue is expected to increase in 2022. For the sales to continue increasing even more Amazon should focus on their marketing strategies which will draw customers attention into purchasing books that have great reviews and many ratings. In this way, the overall revenue of Amazon from the book selling section will increase also [8-10].

Recommendations

Although this analysis was successful in identifying the characteristics of a bestseller, the next step would be to establish how to anticipate how this book will appear in the first few months after its release.

- Continue to analyze the sentiments of user reviews for bestsellers books.
- New books with quickly increasing sales and favorable reviews should be tracked and compared to the bestseller characteristics list.
- Analyze what types of books are becoming popular that year.
- Focus on the marketing strategies to attract new buyers.

References

1. Lee S, Ji H, Kim J, Park E (2021) What books will be your bestseller? A machine learning approach with Amazon Kindle. The Electronic Library 39: 137-151.
2. Leitao L, Amaro S, Henriques C, Fonseca P (2018) Do consumers judge a book by its cover? A study of the factors that influence the purchasing of books. Journal of Retailing and Consumer Services 42: 88-97.
3. Schneider L, Scholten J, Sandor B, Gros C (2021) Charting closed-loop collective cultural decisions: From book best sellers and music downloads to Twitter hashtags and Reddit comments. The European Physical Journal B 94: 161.
4. Yucesoy B, Wang X, Huang J, Barabasi AL (2018) Success in books: A big data approach to bestsellers. EPJ Data Science 7: 1-25.
5. Clay K, Krishnan R, Wolff E, Fernandes D (2003) Retail strategies on the web: Price and non-price competition in the online book industry. The Journal of Industrial Economics 50: 351-367.
6. Bao T, Chang T lung S (2014) Why Amazon uses both the New York Times Best Seller list and Customer Reviews: An empirical study of multiplier effects on product sales from multiple earned media. Decision Support Systems 67: 1-8.
7. Maity SK, Panigrahi A, Mukherjee A (2017) Book reading behavior on good reads can predict the amazon best sellers. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 451-454.
8. Parnell C, Driscoll B (2021) Institutions, platforms, and the production of debut success in Contemporary book culture. Media International Australia 187: 123-138.
9. D'Astous A, Colbert F, Mbarek I (2006) Factors influencing readers' interest in new book releases: An experimental study. Poetics 34: 134-147.
10. Krishnan K, Wan Y (2021) The detection of fake reviews in bestselling books. Journal of Electronic Commerce in Organizations 19: 64-79.