7COM1079-0901-2025 - Team Research and Development Project

Final report title: Investigating the Correlation Between Book Price and Number of Reviews

Group ID: A29

Dataset number: ds239

Prepared by:	Kiran Babu Chukka	24095075
	Nikitha Samala	24065029
	Udaya Venkat Naga Sai Sahithi Nutalapati	24075134
	Deepika Akula	24064766
	Sindhuja Lakshmi Kudikala	24087890

University of Hertfordshire
Hatfield, 2025

Table of Contents

1. Introduction

1.1. Problem statement and research motivation **(100 words)**

In online book market, individual customer engagement has been a major role which includes visibility of the product, cost and the behaviour. On platforms like Amazon, customer decision depends on the number of reviews. However, many people study how online markets work, but they really don't know whether the number of reviews a book has is linked with its price. Most of the research focuses on ratings or sales without checking the number of reviews. Understanding this relationship is important, this study aims to find out if there is a meaningful correlation between price of a book and number of reviews.

1.2. The dataset **(75 words)**

This dataset records the amazon bestsellers with categories, which consists of 550 books from the year between 2009 and 2019. The dataset contains attributes such as book title, author, genre, user ratings, number of reviews, price in USD and release year. These attributes effect on the customer engagement with pricing. The dataset provides analysis of whether the number of reviews has a relationship with book price.

1.3. Research question **(50 words). (50 words)**

This study's major research question is:
"Is there a correlation between book price and the number of reviews?"
Answering this question helps us assess whether books with more customer engagement show uniform pricing variations among Amazon's top-selling titles.

1.4. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

In order to identify whether customer engagement is related to book pricing on Amazon, this study tests the relationship between book price and number of reviews.
Null hypothesis (H0): There is no correlation between book price and number of reviews
Alternative hypothesis (H1): There is a correlation between book price and number of reviews.
By testing these two hypothesis, we can determine whether customer engagement affects pricing in online book platforms.

2. Background research

2.1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**

Studies on digital commerce have often underlined how crucial customer reviews are in influencing market results and consumer behavior. Particularly in competitive environments where consumers depend on social cues to assess quality, Chen et al. (2024) showed that online reviews are essential for developing trust and affecting purchasing probability. Their results point to a general valuation influence from engagement signals.

Pricing behaviour on main online platforms was examined by Li and Wu (2023), who discovered that sellers often react to engagement indicators such ratings and reviews counts—when altering product pricing. This indicates that popular products may have varied pricing patterns because of their greater visibility and demand.

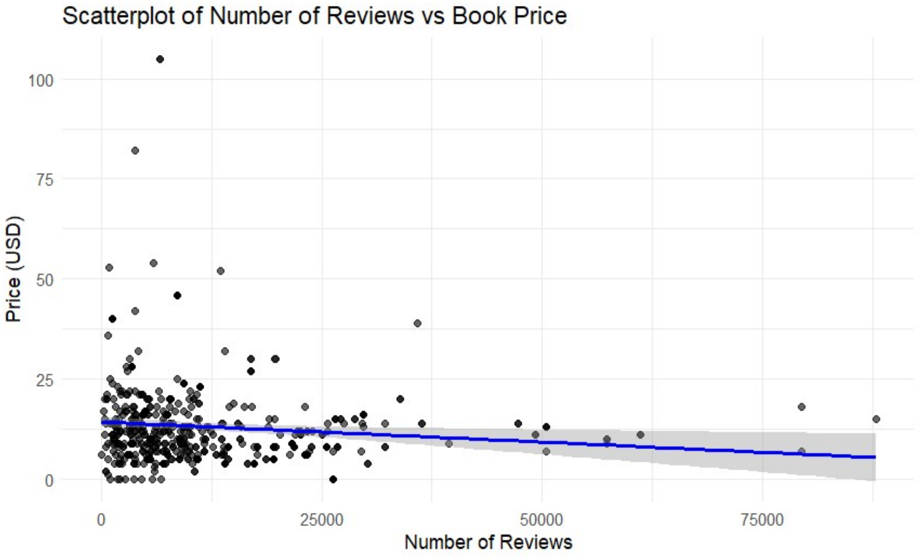
Martinez (2024) investigated how engagement measures help to rank products using algorithms. The study revealed that products with greater review volume usually appear more often in search results, thereby increasing exposure and maybe raising future sales. This visibility may indirectly affect pricing plans as products grow more competitive.

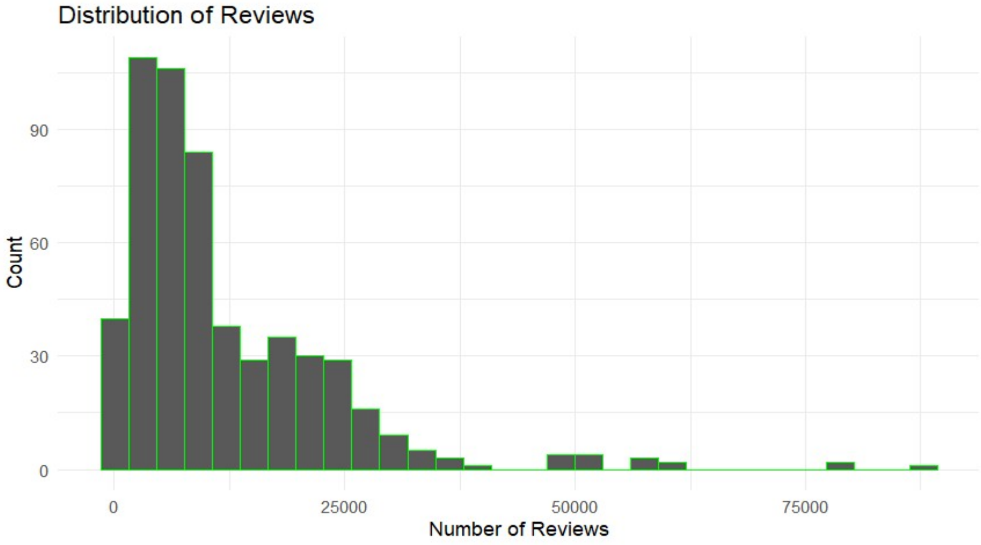
Still, little research investigates if review activity especially inside bestsellers lists straight relates with book pricing. This study tackles this hole by using empirical analysis to find out if book commercial positioning is reflected in engagement indicators.

2.2. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

Although review volume is often seen as a key engagement metric, its link with book price is still unknown. Although bestselling books already have more exposure, it is unknown if this exposure is related to various pricing strategies. Exploring this problem clarifies if books with high ratings are cleverly priced or if lower-priced books naturally draw more readers and hence more reviews. Gaining this knowledge helps to provide more in-depth analysis of how pricing choices, market algorithms, and consumer behaviour interact in digital book markets. This makes the inquiry timely as well as pertinent.

3. Visualisation





3.1. Appropriate graphs for the RQ (**50 words)**

The main graph selected was a scatterplot of Price against Number of Reviews as it best depicts the link between two continuous variables. Additionally included to illustrate how involvement levels are spread inside the dataset was a histogram of review counts. Together, these visuals help the correlation analysis and give context for understanding the outcomes.

3.2. Additional information relating to understanding the data (optional) (**50 words)**

The histogram reveals a pronounced right-skew, with most books getting somewhat few reviews and a little number drawing very great involvement. Though the regression line slopes slightly downward, suggesting a weak negative trend, the scatterplot shows no clear clustering pattern. These remarks support anticipated outcomes before statistical analysis.

3.3 Useful information for the data understanding (**50 words)**

Three observations come out of the images: review activity varies greatly between books; reviews are distributed very unevenly; the scatterplot indicates only a slight downward trend between price and engagement. These results justify the choice to use Pearson correlation to assess if the faint visual trend shows a statistically significant correlation.

4. Analysis

4.1 Statistical test used to test the hypotheses and output (**75 words)**

Since book price and number of reviews are both continuous and fit for a linear relationship test, a Pearson correlation test was chosen. The study resulted in a coefficient of –0.109, which suggests a quite faint negative correlation. Although the effect size is modest, the p-value (0.0104) confirms that this link is statistically significant, therefore unlikely to have happened by chance.

4.2. The null hypothesis is rejected /not rejected based on the p-value (**100 words)**

The null hypothesis is rejected since the p-value of 0.0104 falls short of the 0.05 significance level. Book price and review count have a statistically significant link, this result demonstrates. Though modest, the unfavorable trend indicates that books having more reviews usually sell somewhat less. This trend could point to competitive pricing tactics, greater sales volumes, or algorithmic exposure highlighting cheaper titles. Given the low correlation coefficient, review activity by itself, however, is not a reliable indicator of price; rather, it should be seen as among several variables influencing price.

5 Evaluation – group's experience at 7COM1079

5.1 What went well **(75 words)**

From the start, only three in the group worked well together; regularly helped with analysis, coding, and writing. Regular contact made sure everyone knew the research direction and that chores were finished on time. Using R also improved our technical competence, so we could create dependable statistical results and visualisations. Generally speaking, good organisation and cooperation helped the project grow.

5.2 Points for improvement **(75 words)**

Starting the research sooner would have allowed us more time to investigate other statistical approaches and improve the clarity of our graphics. Some group members needed extra help with R programming, hence having more group coding meetings could have helped everyone understand things better. Earlier management of the writing process would have also assisted guarantee stylistic uniformity throughout the report's sections.

5.3. Group's time management (**50 words)**

Though the project got off to a slow start, once responsibilities were clearly delegated and scheduled meetings started, our time management got better. Setting internal deadlines guaranteed each phase was finished effectively and kept momentum going. Better early planning would have improved our flow more, but on the whole the team kept a decent pace.

5.4. Project's overall judgement (**50 words)**

The project met its goals and demonstrated a robust method for statistical study. Our group carefully examined the data, generated clear and informative graphics, and interpreted the findings sensibly using relevant analysis techniques. Our teamwork and technical abilities were strengthened by the experience, which resulted in a report providing a focused and well-supported response to the research inquiry.

5.5 Comment on the GitHub log output **(50 words)**

Appendix B's GitHub log shows a deliberate progression from data preparation to ultimate analysis. Important commits include plot development, data cleaning, statistical analysis, and report writing. These updates show disciplined team effort and show how the project developed over time, thereby ensuring transparency in the design process.

The three most significant commits during this project are:
1.Commit Message: git commit -m "Add Scatterplot.R for analysis"
Implemented the pearson correlation analysis including scatterplots and histograms to find the correlation between book price and number of reviews.
2. Commit Message: git commit -m readme.txt

In this text file, it contains the output of the null hypothesis which clarifies the significant statistics.

3. Commit message: git commit -m archive (1).zip
This is the dataset of our particular project which contains the bestselling books in amazon. It has attributes like price, number of reviews which affects the customer engagement.

6 Conclusions

6.1 Results explained (**75 words)**

The correlation test showed a small but real negative link between book price and the number of reviews. In other words, books with more reviews usually cost a bit less, though the difference isn't huge. You can see this in the scatterplot—a faint downward slope connects the dots. The histogram, on the other hand, makes it clear that reviews aren't spread out evenly; some books get a ton of attention while others barely get noticed.

6.2 Interpretation of the results (**75 words)**

So, it looks like customer engagement and pricing are linked, but just getting more reviews doesn't really push the price up or down. Sometimes cheaper books sell more, which brings in more reviews, or maybe those bestsellers get a boost from algorithms that favour titles with lots of activity. Still, the connection between price and reviews is pretty weak. Price depends on a bunch of things, and reviews are just one small part of the story.

6.3 Reasons and/or implications for future work, limitations of your study (**50 words)**

Future research could look into more factors like genre, user rating, or publication year. Regression or machine learning methods could provide deeper insight. Since the dataset includes only bestseller books, the findings might not apply to all titles. The uneven distribution of reviews also indicates that exploring non-linear modelling could be valuable.

7. Reference list

- Vasyliuk, A., Matseliukh, Y., Batiuk, T., Luchkevych, M., Shakleina, I., Harbuzynska, H. and Kondratiuk, S. *Intelligent Analysis of Best-Selling Books Statistics on Amazon*. CEUR Workshop Proceedings. This study analyzes Amazon bestseller data including price, number of reviews, and correlation patterns. https://www.scribd.com/document/808025685/amazon-best-selling-books-analysis (Accessed: October 2025)

- Chisomnwa (2025) *Amazon Best Selling Books Analysis*. GitHub repository. Available at: https://github.com/Chisomnwa/Amazon-Best-Selling-Books-Analysis This project explores trends in Amazon bestseller data including price, reviews and genres. https://github.com/Chisomnwa/Amazon-Best-Selling-Books-Analysis (Accessed: 11 December 2025).

- PythonGeeks Team : *Amazon Bestselling Books Analysis Project using Machine Learning*. PythonGeeks.org. Provides insights into Amazon book characteristics and review distributions from a real dataset. https://pythongeeks.org/machine-learning-amazon-book-analysis/ (Accessed: August 2023)

- GitHub – luminati-io (2025) *Amazon Popular Books Dataset*. GitHub repository. A dataset of most reviewed and popular Amazon books with price and reviews. https://github.com/luminati-io/Amazon-popular-books-dataset (Accessed: 11 December 2023)

- Maity, S.K., Panigrahi, A. and Mukherjee, A. *Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers*. arXiv. Examines characteristics (including reviews) that differentiate Amazon bestsellers. https://arxiv.org/abs/1809.07354 (Accessed: 2018)

8. Appendices

A. R code used for analysis and visualisation ***(not included in the word count)***

Analysis.R code with the appropriate statistics to test the hypotheses.

```
# Installing the packages needed for plots and data handling
install.packages("ggplot2")
install.packages("dplyr")

# Loading the libraries needed for plotting and data handling
library(ggplot2)
library(dplyr)

# Reading the dataset from the CSV file
df <- read.csv("bestsellers with categories.csv")

# Quick look at the basic statistics in the dataset
summary(df)

# -------------------------------------------------------
# Scatterplot to check the relationship between reviews and price
# -------------------------------------------------------
scatter_plot <- ggplot(df, aes(x = Reviews, y = Price)) +
  geom_point(alpha = 0.6) +  # adding points with slight transparency
  geom_smooth(method = "lm", se = TRUE, color = "blue") +  # adding a
            trend line
  labs(
    title = "Scatterplot of Number of Reviews vs Book Price",
    x = "Number of Reviews",
    y = "Price (USD)"
  ) +
  theme_minimal()

print(scatter_plot)

# -------------------------------------------------------
# Histogram to see how review counts are distributed
# -------------------------------------------------------
hist_plot <- ggplot(df, aes(x = Reviews)) +
  geom_histogram(bins = 30, color = "green") +
  labs(
    title = "Distribution of Reviews",
    x = "Number of Reviews",
    y = "Count"
  ) +
  theme_minimal()

print(hist_plot)

# -------------------------------------------------------
# Running Pearson correlation test between reviews and price
# -------------------------------------------------------
cor_test <- cor.test(df$Reviews, df$Price, method = "pearson")

# Showing the test result
print(cor_test)

# -------------------------------------------------------
# Printing a simple conclusion to appear in the log
# -------------------------------------------------------
if (cor_test$p.value < 0.05) {
  print("Null hypothesis rejected: There is a significant correlation between
            book price and number of reviews.")
} else {
  print("Null hypothesis not rejected: There is no significant correlation
            between book price and number of reviews.")
}
```

```
###output :

Pearson's product-moment correlation

data:  df$Reviews and df$Price
t = -2.5713, df = 548, p-value = 0.0104
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.19104436 -0.02581111
sample estimates:
       cor
-0.1091819
```

B. GitHub log output.

```
commit b9725efab6f14fcaf54f53328c8d8b2a8cde7cf5 (HEAD -> main, origin/main,
origin/HEAD)
Merge: 61b9744 be5fe5b
Author: nikitha625 <ns24adl@herts.ac.uk>
Date:   Thu Dec 11 15:12:19 2025 +0000

    Merge branch 'main' of https://github.com/nikitha625/Groupproject

commit 61b97446ec166fdcd1564d73b1815b477a207a8e
Author: nikitha625 <ns24adl@herts.ac.uk>
Date:   Thu Dec 11 15:10:28 2025 +0000

    Rename archive (5).zip to bestsellers_with_categories.zip

commit be5fe5b1e57be8046b055dcf135adab394f91dfd
Author: nikitha625 <ns24adl@herts.ac.uk>
Date:   Thu Dec 11 12:55:56 2025 +0000

    Update readme.txt

commit d5f59644e735509ff4823c267fb57b40317813f3
Author: rxkiran <kc24abz@herts.ac.uk>
Date:   Wed Dec 10 21:14:36 2025 +0530

    Add files via upload

commit 85fc89832a887027665db35f531f6989391d54c8
Author: rxkiran <kc24abz@herts.ac.uk>
Date:   Sat Dec 6 03:50:18 2025 +0530
```