College of Computer and Information Sciences

Department of Information Technology

IT 362 : Principles of Data Science

2st Semester 1446 H

Books Sales Trend

Table of Contents

1. Introduction: 4
1. Data Analysis Workflow 4
2. Our Data: 4
   2 .1 Data Sources: 5
   2 .2 Data Collection: 5
   2 .3 Screenshots: Samples of Semi-Structured and Structured Data. 6
   2 .4 Data Type: 7
   2 .5 Evaluation of Potential Biases in the Data: 8
3. Objectives: 9
4. Method: 9
5. Data Cleaning and Prior Preprocessing 11
   5 .1 Handling Missing Values 11
   5 .2 Cleaning and Formatting Numerical Columns 11
   5 .3 Handling Duplicate Rows and Data Aggregation 11
   5 .4 Genre Cleaning and Standardization 11
   5 .5 Handling the " Author " Column 12
   5 .6 " Book Type " Column 12
   5 .7 Image Processing and Book Cover Classification 12
      5.7.1 Image Processing Using Selenium : Why Did We Use Selenium? 12
      5.7.2 Steps for Image Collection 12
      5.7.3 Extracting Visual Features: Why Extract Visual Features? 13
      5.7.4 Selected Features 13
      5.7.5 Steps for Feature Extraction 13
      5.7.6 Classifying Book Covers : Why Classify Book Covers? 14
      5.7.7 Classification Methodology 14
      5.7.8 Findings 15
6. Exploratory Data Analysis (EDA) 16
   6 .1 EDA: Primary Data Investigation 16
      6.1.1 Statistical measure: 16
      6.1.2 Feature distribution: 17
      6.1.3 Feature relationships: 18
   6 .2 EDA: Secondary Data Investigation 20

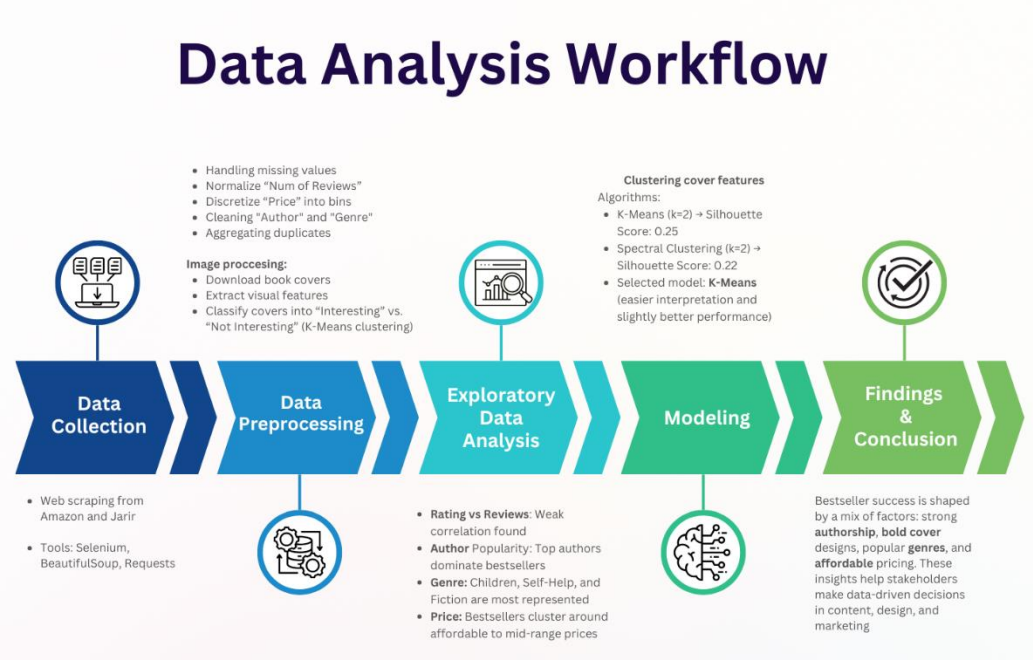
6.2.1 Statistical measure: 20
6.2.2 Feature distribution: 21
6.2.3 Feature relationships: 22
7. Comparison Between Primary and Secondary Data 24
7 .1 Overview of Data Sources 24
7 .2 Key Discrepancies & Alignments 25
7 .3 Interpretation of Findings 26
8. Summary of New Insights and Hypotheses 27
9. Data Preprocessing 27
9 .1 Handling Duplicate Rows and Data Aggregation 27
9 .2 Normalization of 'Num of Reviews' 28
9 .3 Discretization of 'Price' 28
10. Modeling Task : 28
11. Results and Discussion: 32
12. Conclusion and Future Work: 38
13. Challenges: 39
14. Recommendations for Mitigating Challenges: 40
15. References 42

1. Introduction:

With the rapid expansion of the book market and the diversity of genres, there is a growing need to understand the factors that contribute to books being classified as bestsellers and to analyze the patterns associated with them. This project aims to study book data, including the number of reviews and the books that is listed as a bestseller, to uncover key trends such as the most in-demand genres and the factors that attract readers and increase a book's popularity. This analysis is expected to provide valuable insights that can help publishers and authors enhance their marketing strategies and boost the success of their books.

1. Data Analysis Workflow

The mentioned figure gives overview of the data analysis workflow, illustrating the key phases from data collection to final findings.



2. Our Data:

We primarily applied web scraping on Amazon's and Jarir's bestseller books to collect our data. As for our dataset, we collected the top 100 – 200 bestsellers from each website across different genres.

2 .1 Data Sources:

We collected our data by web scraping from the following online stores:

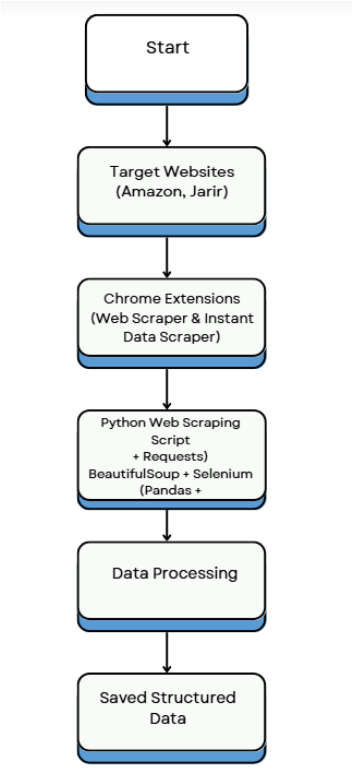
- **Amazon[4]** is a global leader in e-commerce, offering a wide range of products and services. Its online marketplace features an extensive collection of books spanning various genres, formats, and price ranges. This makes Amazon an invaluable resource for analyzing trends in book sales and understanding the factors contributing to their success.
- **Jarir Bookstore[5][6]** is a leading retailer in the Middle East, specializing in books, electronics, office supplies, and educational materials. Its online and physical stores feature a diverse collection of books across multiple genres, formats, and price ranges. Jarir's strong regional presence and reputation make it an essential resource for analyzing book sales trends and understanding the factors contributing to their popularity in the Middle Eastern market.

These sources were chosen because they represent a diverse range of books, have a large and diverse audience, and provide relatively complete data. By focusing on bestseller lists, we aim to study the factors that contribute to a book's success in these markets.

2 .2 Data Collection:

In this project, web scraping was used to collect data on the top 100 bestseller books from various sources. Two Chrome extensions assisted in data extraction: Web Scraper - Free Web Scraping [7], which helped bypass server restrictions on Amazon, and Instant Data Scraper - Free Web Scraping [8], which was used to extract bestseller book URLs from the Jarir bookstore website.

For the web scraping script, several Python libraries and tools were utilized. requests and BeautifulSoup handled static HTML extraction, while Selenium was used to automate browser interactions and scrape dynamically loaded content. pandas facilitated data processing and storage, and time was used to introduce delays between requests. The script incorporated standard methods to fetch web pages, locate elements, and extract relevant data efficiently.



This figure illustrates the data collection workflow. Initially, bestseller book information was gathered from Amazon and Jarir websites. Chrome extensions (Web Scraper and Instant Data Scraper) assisted in extracting preliminary data. A custom Python script utilizing Requests, BeautifulSoup, Selenium, and Pandas was developed to scrape, process, and organize the data into a structured format ready for analysis.

2 .3 Screenshots: Samples of Semi-Structured and Structured Data

To demonstrate the transformation process, the following figures present samples of the dataset before and after preprocessing. The initial collected data required significant cleaning to correct missing values, inconsistent formats, and partially unorganized fields, ultimately producing a fully structured dataset ready for analysis.

Note: Although the data was extracted in a tabular format through web scraping, it was considered semi-structured due to the presence of missing values and inconsistent formatting in various fields. Significant preprocessing was necessary to transform it into a fully structured dataset ready for analysis.

- **Sample of Raw Semi-Structured Data**

	Title	Price	Rating	Num Of Reviews	Author	Book Type	Genre	Cover Image
0	كتاب التحصيلي علمي 46-47 (2025)	SAR 98.00	4.3 out of 5 stars	10	Nasser bin Abdulaziz Al-Abdulkarim	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
1	El Sharq library المعاصر 9 تأسيس كمي 2/1 ورقي ...	SAR 107.58	4.5 out of 5 stars	230	صياد الجزيري	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
2	My First Library : Boxset Of 10 Board Books Fo...	SAR 59.00	4.6 out of 5 stars	80,714	Wonder House Books	Board book	Null	https://images-eu.ssl-images-amazon.com/images...
3	فاتتني صلاة	SAR 26.00	4.7 out of 5 stars	301	اسلام جمال	Unknown Binding	Null	https://images-eu.ssl-images-amazon.com/images...
4	The Power of your Subconscious Mind	SAR 35.00	4.5 out of 5 stars	14,131	Joseph Murphy	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
5	Coloriages mystères Disney Princesses: Colorie...	SAR 109.10	4.7 out of 5 stars	5,880	Jérémy Mariez	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
6	101 Unicorn Colouring Book: Fun Activity Colou...	SAR 21.62	4.6 out of 5 stars	2,321	Wonder House Books	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
7	White Nights	SAR 19.00	4.6 out of 5 stars	1,583	Fyodor Dostoyevsky	Mass Market Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
8	Null	SAR 65.00	4.7 out of 5 stars	12,588	4.7 out of 5 stars 12,588	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...
9	My First 365 Coloring Book: Jumbo Coloring Boo...	SAR 29.00	4.5 out of 5 stars	1,625	Wonder House Books	Paperback	Null	https://images-eu.ssl-images-amazon.com/images...

This figure shows a sample of the raw collected data before preprocessing. The data contains missing values in the Genre column, mixed textual and numerical information in the Price and Rating fields, and some inconsistencies that required cleaning and restructuring.

- **Sample of Final Structured Data**

	Title	Price	Rating	Num Of Reviews	Author	Book Type	Genre	Cover Image
0	يقولون الأولين	3	5.0	0.00905	Walid Al-Anazi	Paperback	Fiction Genres	https://www.jarir.com/cdn-cgi/image/fit=contai...
1	صفحة واحدة في المرة رفيق مبدع يومي	3	4.5	0.00459	Adam J. Kertz.	Paperback	Self-Help & Personal Development	https://www.jarir.com/cdn-cgi/image/fit=contai...
2	المعاصر 9 تأسيس كمي ورقمي ومحوسب مع بنك المحوس...	4	4.7	0.00360	Imad Al-Zarri	Paperback	Education & Reference	https://www.jarir.com/cdn-cgi/image/fit=contai...
3	الأب الغني الأب الفقير ما يعلمه الأثرياء ولا ي...	4	5.0	0.00298	Robert T. Kiusaki	Paperback	Business & Finance	https://www.jarir.com/cdn-cgi/image/fit=contai...
4	مصحف القيام مقاس 25×35	4	5.0	0.00248	Fatima. عاتف مقوى فني		Religion & Spirituality	https://www.jarir.com/cdn-cgi/image/fit=contai...
...	...	...	...	...	...	...	...	...
334	The Ballad of Never After: the stunning sequel...	3	4.8	0.32477	Stephanie Garbe r	Paperback	Science Fiction & Fantasy	https://images-eu.ssl-images-amazon.com/images...
335	No Longer Human	3	4.7	1.69343	Osamu Dazai	Paperback	Historical Fiction	https://images-eu.ssl-images-amazon.com/images...
336	The Coming Wave: Technology, Power, and the Tw...	3	4.4	0.10962	Mustafa Suleyman	Paperback	Technology & Digital Media	https://images-eu.ssl-images-amazon.com/images...
337	IKIGAI	4	4.6	5.22402	Héctor García	Hardcover	Self-Help & Personal Development	https://images-eu.ssl-images-amazon.com/images...
338	a good girl's guide to murder	3	4.8	1.46067	Holly Jackson	Paperback	Fiction Genres	https://images-eu.ssl-images-amazon.com/images...

This figure shows a sample of the final structured dataset after preprocessing. Missing values were addressed, textual and numerical inconsistencies were corrected, and the data was organized into clean, analyzable fields suitable for modeling and analysis.

2 .4 Data Type:

To study the factors influencing bestseller books, we identified key attributes that are likely to have a significant impact on a book's popularity. After reviewing related studies, research papers, and articles, we referenced the following sources to guide our attribute selection [1] [2] [3] .

- Each **row** represents a book which we will discuss some features about it which represents the **columns.** We have collected 341 books so far and each one of them have 8 features.

Feature	Type of data	Measurement level	Description
Title	Qualitative	Nominal	The name of the book.
Rating	Quantitative	Interval	Average rating (1-5).
Price	Quantitative	Ratio	Price of the book in the local currency.
Num Of reviews	Quantitative	Ratio	The number of text reviews for the book.
Book type	Qualitative	Nominal	Type of the book (Paper or E-book).
Author	Qualitative	Nominal	Name of the author.
Cover Image	Qualitative	Nominal	URL for the book's cover image .
Genre	Qualitative	Nominal	Defines book's theme or style.

2 .5 Evaluation of Potential Biases in the Data:

• Representation:

The data focuses on popular books in Saudi Arabia, most of which are in Arabic. This aligns well with the project's objective, but it may not sufficiently cover all categories, such as non-Arabic books, which are underrepresented in our dataset, or important books that are not listed as bestsellers.

• Measurement Bias:

The bestselling books on Amazon and Jarir can be influenced by advertisements, book availability, or even the rating systems used.

• Historical Biases:

The data may reflect long-standing biases in the publishing market, such as a focus on traditional authors or topics at the expense of newer authors or contemporary themes. Certain groups, like young authors or female writers, might be underrepresented due to historical consumption patterns that favor specific types of authors or subjects.

3. Objectives:

To achieve our main goal of understanding the factors that contribute to books being classified as bestsellers and analyzing the patterns associated with them, we conducted a thorough review of related studies, research papers, and articles [1][2][3]. Building on the insights gained from this literature review, we formulated the following research questions:

- How do ratings and the number of reviews vary among bestsellers?
- Are certain authors more likely to have their books become bestsellers?
- What visual design patterns are consistently associated with bestseller book covers?
- What genres are most represented among bestsellers?
- What is the relationship between price and bestseller books?

4. Method:

After collecting and processing the data, we plan to answer our key questions to uncover patterns among bestsellers and identify factors contributing to their success.

1. **How do ratings and the number of reviews vary among bestsellers?**

   We will analyze the distribution of ratings and review counts using histograms, scatter plots or heatmap. Correlation analysis will help determine if higher-rated books tend to receive more reviews, highlighting engagement trends among bestsellers and analyze patterns in their genres and ratings.

2. **Are certain authors more likely to have their books become bestsellers?**

By counting each author's appearances on the bestseller list and visualizing the data with bar charts or network graphs, we can identify authors with multiple bestsellers.

3. **What visual design patterns are consistently associated with bestseller book covers?**

   We applied image processing techniques to analyze design elements such as color schemes, typography, and visual complexity. Using clustering algorithms, we assessed whether certain visual features are more common among bestseller book covers.

4. **What genres are most represented among bestsellers?**

   We will categorize books by genre and use visualizations such as bar charts, pie charts or treemaps to highlight the most common bestseller categories. Comparing genre trends with ratings and reviews will reveal which genres attract the most reader engagement.

5. **What is the relationship between price and bestseller books?**

   We will examine price distributions with histograms and box plots to understand bestseller pricing trends. Correlation analysis will help determine if price influences popularity or if specific price ranges dominate the bestseller market.

Using previous plans to answer main problem

Our analysis of ratings, reviews, authors, cover design, genres, and pricing provides key insights into what drives a book's success. Strong reader engagement, frequent bestsellers from established authors, and popular genres all contribute to market trends. Pricing strategies also play a role, revealing optimal price ranges for bestsellers. These insights help publishers and authors refine their marketing strategies. Using predictive models such as regression analysis, we can further explore the impact of different factors on a book's success.

5. Data Cleaning and Prior Preprocessing

Objective: to prepare the data for Exploratory Data Analysis (EDA) and ensure meaningful insights, we apply the following techniques to both primary and secondary data separately.

5 .1 Handling Missing Values

To address missing values in the dataset, we implemented three specific strategies. For Unknown Authors, we retained the label "Unknown" as it indicates that the website did not specify the author's name, rather than representing a true missing value. Ratings and Reviews with null values were replaced with zero to signify that no ratings or reviews were provided. Rows with Missing Author Names (completely null) were dropped from the dataset, as they lacked essential information.

5 .2 Cleaning and Formatting Numerical Columns

To clean and standardize the dataset, we focused on three key columns: Price, Rating, and Num of Reviews. Using regular expressions, we extracted the numeric values from these columns to ensure only relevant numbers were captured. This process ensured that the data remained consistent, accurate, and ready for further analysis.

5 .3 Handling Duplicate Rows and Data Aggregation

To address duplicate rows in the dataset, we first cleaned the Title column by converting the text to lowercase and removing extra spaces. We then identified rows that had the same title.

5 .4 Genre Cleaning and Standardization

To clean and standardize the dataset, we focused on the Genre column. First, we removed duplicate commas and extra whitespace to ensure proper formatting. Next, unnecessary genre labels such as "Best Sellers" and "New Arrivals" were removed from both the beginning and end of the genre string.

We then mapped various genre labels to standardized categories such as "Fiction Genres" and "Non-Fiction Genres" using a predefined dictionary. This process ensured that labels, making the data ready for further analysis.

5 .5 Handling the " Author " Column

We encountered an issue with the 'Author' column, where the data contained a mix of Arabic and English names. To standardize this and ensure consistency, we translated the Arabic text into English using argostranslate.

5 .6 " Book Type " Column

Regarding the 'Book Type' column, we chose not to perform any preprocessing at this stage. While we recognize that it contains data in multiple languages and may have some inconsistencies, we decided to leave it as is for now. If, in the future, we determine that preprocessing is necessary, we will address it. Otherwise, we may choose to drop the column altogether.

5 .7 Image Processing and Book Cover Classification

Analyzing book covers is essential to understanding the impact of visual elements on a book's appeal. In this project, we collected book cover images and analyzed them by extracting specific visual features that help classify them into different categories.

5.7.1 Image Processing Using Selenium : Why Did We Use Selenium?

At first, we tried to access book cover images directly from the database links. However, we discovered that many websites protect their content from direct downloads, making it impossible to retrieve the images using traditional methods. To solve this issue, we utilized Selenium, a tool that simulates user interactions on web pages, such as opening pages and interacting with elements, allowing us to extract images as if we were manually browsing the website.

5.7.2 Steps for Image Collection

1. **Setting Up Selenium WebDriver**: We configured Selenium to run in headless mode, enabling it to browse websites automatically without displaying a browser window.
2. **Extracting Image URLs from the Database**: We gathered and validated the URLs of book cover images.

3. **Downloading Images Using Selenium**: We used Selenium to visit each link, locate the image element, and capture a screenshot of the book cover.
4. **Storing Images Locally**: The images were saved in a dedicated folder for further analysis.

Thanks to Selenium, we overcame technical restrictions that prevented direct downloads and successfully collected a comprehensive dataset of book cover images. These images will be used in the next steps to extract visual features and analyze their impact on book cover appeal.

5.7.3 Extracting Visual Features: Why Extract Visual Features?

To analyze the appeal of book covers, we extracted key visual features that influence human perception and design aesthetics. Our selection of these features is supported by research in visual psychology and marketing

Demonstrating those elements such as contrast, edge complexity, and color properties play a crucial role in capturing attention and influencing consumer behavior [9].

5.7.4 Selected Features

We identified four primary visual features to analyze:

1. **Contrast** – Measures the difference between light and dark areas in an image, affecting readability and attention.
2. **Edge Complexity** – Evaluates the level of detail in a cover by detecting the number of edges, which indicates design intricacy.
3. **Color Variance** – Assesses the diversity of colors used in the design, as high variance can create a dynamic and visually appealing effect.
4. **Color Saturation** – Determines the intensity of colors, where higher saturation levels often lead to more engaging and vibrant covers.

5.7.5 Steps for Feature Extraction

1. **Convert Image to Grayscale for Contrast Analysis**: Grayscale images help in computing contrast by analyzing brightness differences.

2. **Apply Edge Detection for Complexity Measurement**: The Canny edge detection algorithm is used to quantify the amount of detail present.
3. **Compute Color Variance and Saturation**: RGB values are analyzed to determine color diversity, while HSV conversion is used for saturation analysis.
4. **Store Extracted Features in a Dataset**: Each extracted feature is recorded and associated with its respective book cover for further classification.

The extracted visual features provide a structured way to evaluate book covers based on their design elements. These features will be used in the next step to classify covers into categories of "interesting" or "not interesting," aiding in objective cover assessment. Our approach is supported by previous research in visual aesthetics, marketing, and cognitive psychology, ensuring that the classification aligns with human perception and industry standards.

5.7.6 Classifying Book Covers: Why Classify Book Covers?

After extracting the visual features, the next step is to classify book covers into "interesting" or "not interesting" categories. Classification helps in understanding which design elements contribute to an engaging book cover. By utilizing machine learning techniques, we can create an objective approach to assessing book cover appeal based on quantifiable features.

5.7.7 Classification Methodology

To classify book covers, we applied **K-Means Clustering**, a widely used unsupervised machine learning algorithm. K-Means groups data into clusters based on feature similarities. Since we aimed to distinguish between two categories, we set the number of clusters to two.

The clustering process involved the following steps:

1. **Feature Normalization** – Before applying K-Means, we normalized the extracted visual features to ensure equal weighting across all attributes.
2. **Applying K-Means Algorithm** – We initialized the K-Means algorithm with two clusters and trained it on our dataset of book cover features.

3. **Determining the "Interesting" Cluster** – After clustering, we analyzed the mean values of each cluster to identify which group contained the more visually appealing book covers.
4. **Assigning Labels** – Covers that belonged to the "interesting" cluster were labeled as **1 (Interesting)**, while those in the other cluster were labeled as **0 (Not Interesting).**

By leveraging K-Means Clustering, we successfully categorized book covers based on their visual appeal. This classification provides a data-driven approach to evaluating book cover effectiveness, allowing for insights into which design elements contribute to a more engaging and marketable book cover. The results from this classification can be further refined through supervised learning models in future research to improve accuracy and interpretability.

5.7.8 Findings

We calculated the correlation between book cover features and their classification as "Interesting" to determine the most influential factors in cover appeal. The results indicate that color characteristics play a significant role in determining a book cover's attractiveness. Mean green (-0.77), mean blue (-0.75), and mean red (-0.68) show strong negative correlations, suggesting that darker or muted colors may reduce a cover's visual appeal. In contrast, contrast (0.56), color variance (0.68), and color saturation (0.66) exhibit positive correlations, indicating that high contrast and vibrant colors enhance visual appeal and increase the likelihood of a book being perceived as interesting.

Additionally, edge complexity (0.22) and sharpness (0.02) show minimal influence, meaning that intricate details and image sharpness do not significantly impact a book cover's attractiveness. Based on these findings, it is recommended that book cover designs prioritize bold colors, increased contrast, and diverse color schemes to enhance their appeal and engage readers effectively.

6. Exploratory Data Analysis (EDA)

Objective: Investigate key factors influencing book popularity by analyzing ratings, reviews, bestseller attributes, pricing, genre distribution, and author impact.

Library used:

- Matplotlib: For basic plotting functionality and figure layout management.
- Seaborn: For creating enhanced statistical visualizations such as raincloud plots, bar charts, and violin plots.
- Plotly: For interactive distribution plots to compare rating distributions dynamically.

6 .1 EDA: Primary Data Investigation

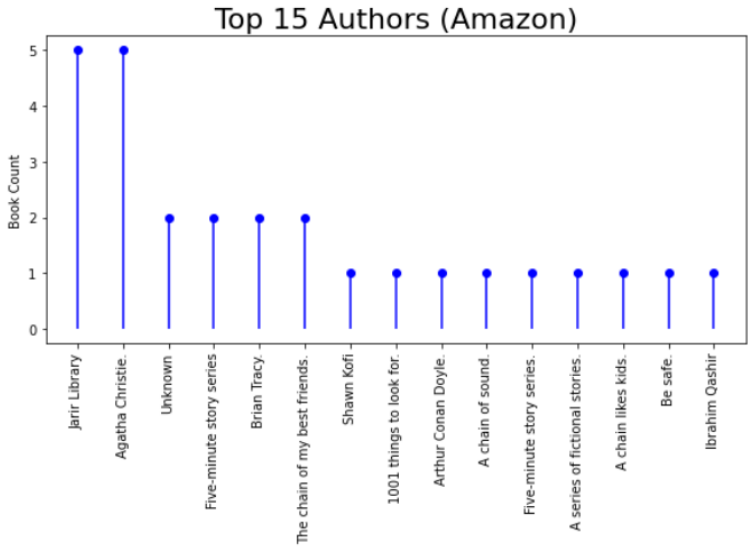
6.1.1 Statistical measure:

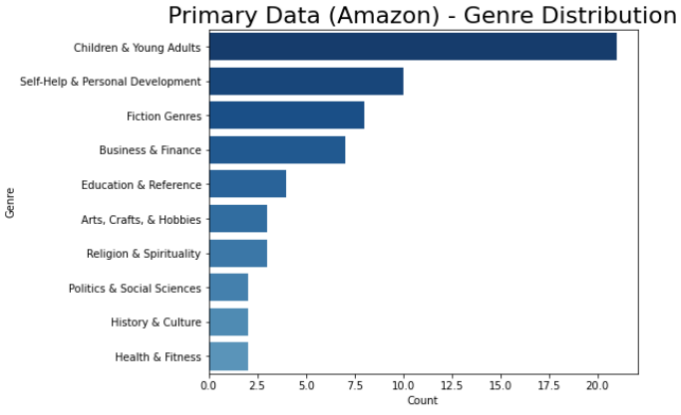
	Amazon Statistics (Primary)			
	Price	**Rating**	**Num Of Reviews**	**Interesting**
count	79.000000	79.000000	79.000000	79.000000
mean	63.448734	4.578481	9077.417722	0.177215
std	37.430640	0.540102	16905.904455	0.384291
min	9.000000	0.000000	0.000000	0.000000
25%	35.735000	4.550000	394.000000	0.000000
50%	55.000000	4.600000	2535.000000	0.000000
75%	84.290000	4.700000	9284.500000	0.000000
max	228.470000	4.900000	80641.000000	1.000000

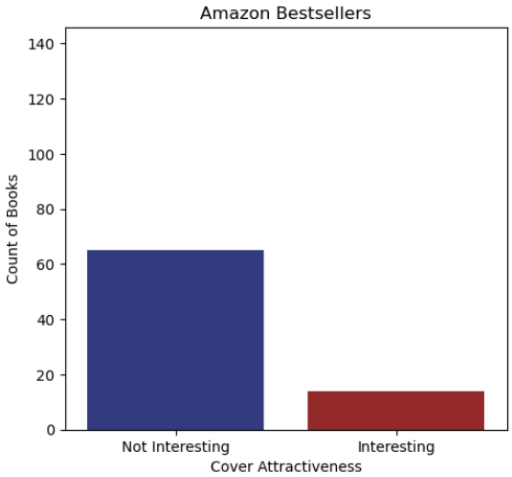
To gain an overview of the statistical properties of numerical attributes, we used the describe() function. The analysis revealed that Amazon products have a mean rating of 4.57, indicating generally high customer satisfaction. Additionally, the mean number of reviews is 9,077, suggesting a large and highly engaged customer base. In terms of pricing, the mean price of Amazon products is 63.44 SR, indicating that Amazon's product offerings tend to be in the mid-to-high price range within the dataset. Furthermore, only 17.7% of

Amazon products are marked as "interesting" (mean = 0.177), which implies that cover attractiveness may not be a primary driver of customer interest or sales.

6.1.2 Feature distribution:







Analysis and Insights:

- Genre Distribution

Prominent Genres: The "Children & Young Adults" genre dominates with the highest count, followed by "Fiction Genres," "Self-Help & Personal Development," and other categories such as "Arts, Crafts, & Hobbies" and "Psychology & Philosophy." The count distribution in Amazon data is fairly wide across genres, with a noticeable peak for the top genres (especially "Children & Young Adults"), tapering off as the genres become more niche.

- Author Distribution

Greer's Library (Jarir Bookstore) and Agatha Christie have the highest number of books (~5 each). Most other authors have only 1-2 books, indicating a broad diversity of authors with relatively small individual contributions. Interpretation Amazon's catalog seems to contain a mix of well-known and niche authors, providing a wider selection.

- Cover Attractiveness Distribution

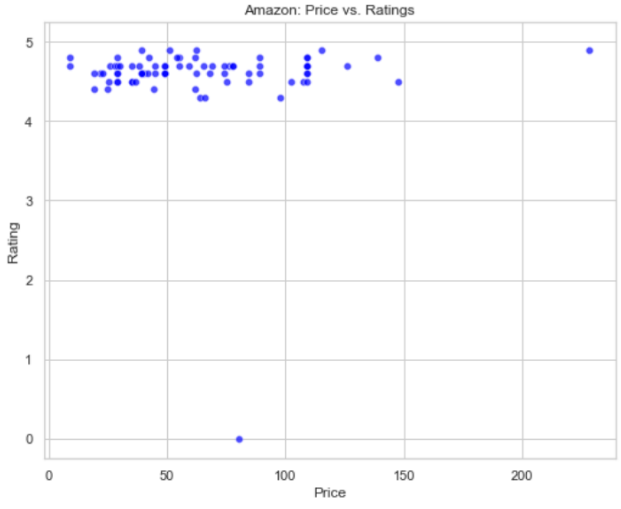
A large majority of Amazon bestsellers fall under the "Not Interesting" category, meaning that cover attractiveness does not strongly correlate with bestseller status in this dataset. Only a small fraction of books are labeled as "Interesting," suggesting that other factors such as reviews, ratings, or author popularity may have a more significant influence on sales. Since Amazon operates in a highly competitive online environment, cover design alone may not be a determining factor in a book's success.

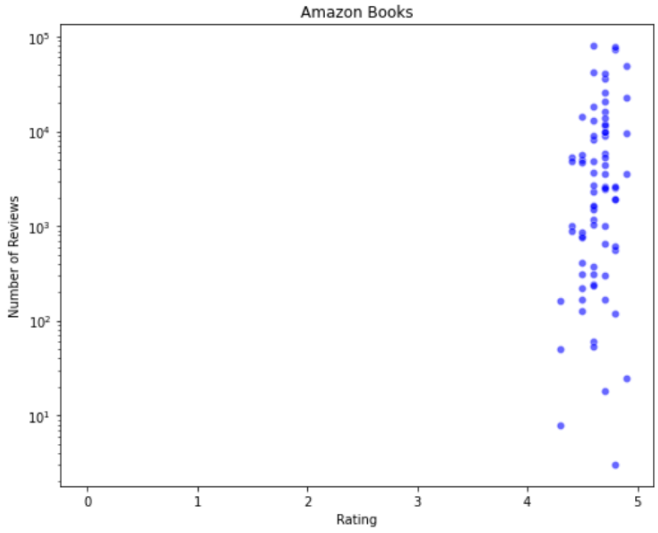
6.1.3 Feature relationships:

Amazon Correlation

	Price	Rating	Num Of Reviews	Interesting
Price	1.000000	-0.006813	-0.095441	-0.122198
Rating	-0.006813	1.000000	0.127249	0.024786
Num Of Reviews	-0.095441	0.127249	1.000000	0.033640
Interesting	-0.122198	0.024786	0.033640	1.000000

For Amazon products, the analysis reveals several key correlations. First, there is a **weak positive correlation (0.127)** between **Rating** and **Number of Reviews**, suggesting that higher-rated products tend to receive slightly more reviews. Second, there is a **weak negative correlation (-0.0068)** between **Price** and **Rating**, indicating that higher-priced items might have slightly lower ratings. Finally, there is a **negative correlation (-0.095)** between **Price** and **Number of Reviews**, meaning that higher-priced items tend to receive fewer reviews. These insights highlight the nuanced relationships between pricing, customer satisfaction, and engagement in the Amazon dataset.





Analysis and Insights:

Books with high ratings (4-5) tend to have significantly more reviews, with the distribution being highly skewed as a few books have an extremely high number of reviews. On the other hand, books with ratings below 4 have considerably fewer reviews, suggesting that bestsellers generally receive higher ratings.

When examining the relationship between **price and ratings**, there is no strong correlation, as highly rated books exist across various price points. Most best-rated books (close to 5) are spread across different price ranges, indicating that price does not heavily influence user ratings. Additionally, a majority of books are clustered around mid-range prices, with a few high-priced books maintaining strong ratings.

6 .2 EDA: Secondary Data Investigation

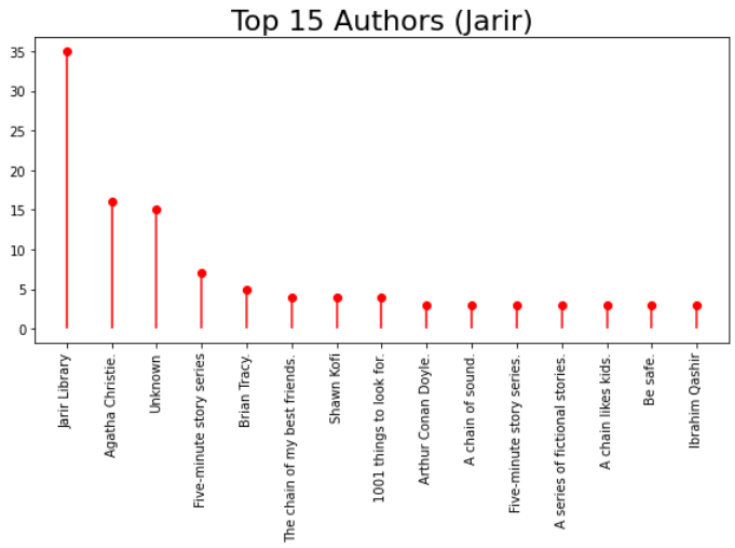
6.2.1 Statistical measure:

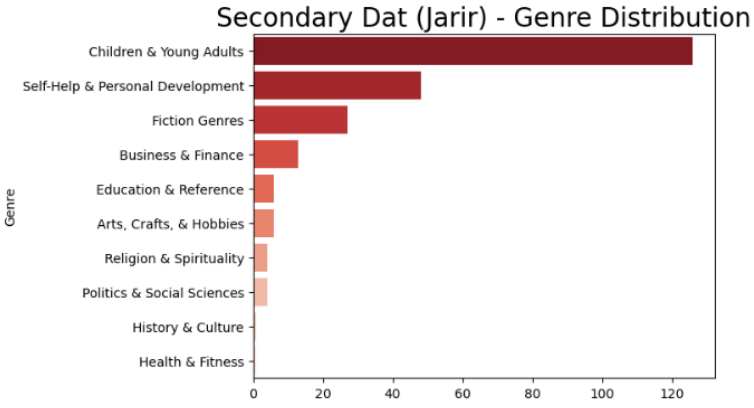
Jarir Statistics (Secondary)	Price	Rating	Num Of Reviews	Interesting
count	236.000000	236.000000	236.000000	236.000000
mean	41.690678	2.904237	3.250000	0.588983
std	21.208431	2.385966	8.847707	0.493064
min	2.000000	0.000000	0.000000	0.000000
25%	26.500000	0.000000	0.000000	0.000000
50%	37.000000	4.600000	1.000000	1.000000
75%	49.000000	5.000000	2.250000	1.000000
max	145.000000	5.000000	91.000000	1.000000

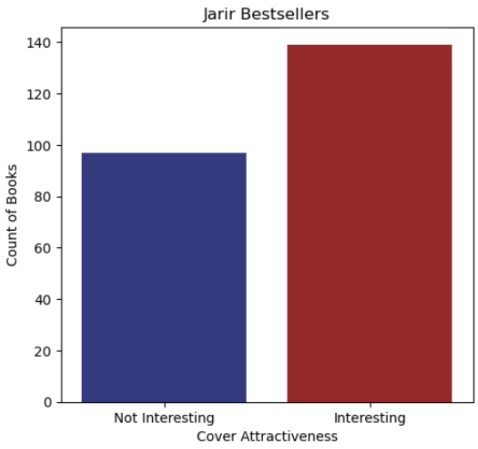
To gain a brief overview of the statistical properties of numerical attributes, we used the describe() function. The analysis revealed that Jarir products have a mean rating of 2.90, indicating moderate to low customer satisfaction. Additionally, the mean number of reviews is 3.25, suggesting a smaller customer base. In terms of pricing, the mean price of Jarir products is 41.69 SAR, which may suggest a focus on budget-friendly or mid-range products. Furthermore, approximately 58.9% of Jarir products are marked as "interesting"

(mean = 0.589), indicating that cover attractiveness may play a significant role in driving customer interest.

6.2.2 Feature distribution:







Analysis and Insights:

- Genre Distribution

In the Jarir dataset, the "Children & Young Adults" genre stands out as the most prominent, with a significant lead in book counts. However, there is a steep drop-off in counts for genres following the leading category. Genres such as "Fiction Genres" and "Self-Help & Personal Development" have relatively lower representation, indicating that Jarir's catalog is heavily focused on a specific set of genres. This contrasts with Amazon's

more evenly distributed genre representation, highlighting Jarir's emphasis on "Children & Young Adults" as a key category.

- Author Distribution

Greer's Library (Jarir Bookstore) is the dominant author in the dataset, with approximately 25 books, followed by Agatha Christie, who has a strong presence with around 15 books. The counts for other authors show a gradual decline, indicating a long tail of less prominent contributors. Notably, the presence of "Unknown" as one of the top-ranked authors is unusual and could suggest potential data issues or an interesting pattern worth further investigation.

- Cover Attractiveness Distribution

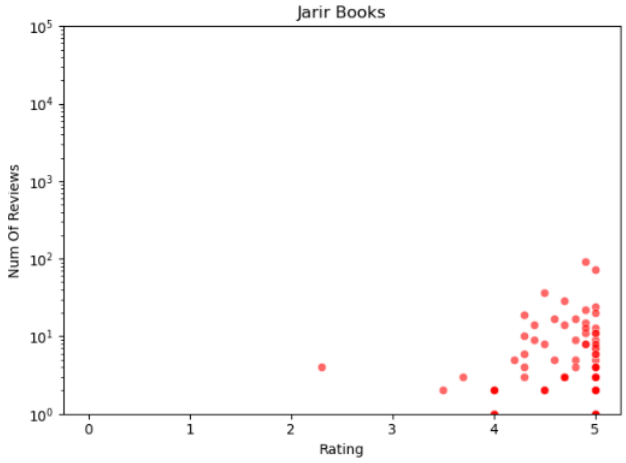
Jarir's bestsellers exhibit a more balanced distribution between "Interesting" and "Not Interesting" covers compared to Amazon. A higher proportion of books at Jarir are labeled as "Interesting," suggesting that cover design may play a more significant role in driving consumer interest. This trend could be influenced by the in-store shopping behavior of Jarir's customers, who physically browse books, making visual appeal a more impactful factor in their purchasing decisions.

6.2.3 Feature relationships:

Jarir Correlation (Secondary)	Price	Rating	Num Of Reviews	Interesting
Price	1.000000	0.227666	0.191403	0.010172
Rating	0.227666	1.000000	0.293685	-0.084601
Num Of Reviews	0.191403	0.293685	1.000000	-0.003658
Interesting	0.010172	-0.084601	-0.003658	1.000000

For Jarir products, the analysis reveals several key correlations. First, there is a **moderate positive correlation (0.294)** between **Rating** and **Number of Reviews**, indicating that higher-rated products are more likely to receive more reviews compared to Amazon. Second, there is a **weak positive correlation (0.227)** between **Price** and **Rating**, suggesting that higher-priced items might have slightly higher ratings. Finally, there is a **weak positive correlation (0.191)** between **Price** and **Number of Reviews**, meaning

that higher-priced items tend to receive slightly more reviews. These insights highlight the relationships between pricing, customer satisfaction, and engagement in the Jarir dataset.



Analysis and Insights:

- Price & Ratings

A significant number of books in Jarir's dataset have low ratings across all price ranges, showing that price does not strongly influence customer ratings. Unlike Amazon, where higher-priced books tend to receive higher ratings, Jarir's data lacks a clear pattern. Some higher-priced books have decent ratings, but lower-priced books do not necessarily have worse ratings, suggesting other factors drive customer satisfaction.

Contradictions & Possible Explanations

- **User Engagement:** Jarir customers may not rate books as frequently as Amazon users.
- **Rating Culture:** A bimodal pattern suggests users either rate highly or not at all.
- **Market Differences:** Amazon has global price competition, while Jarir serves a regional market.
- **Data Gaps:** Many books have near-zero ratings, possibly due to missing data.

- Rating & Review

Many books on Jarir's platform have very few reviews, especially those with low ratings. Unlike Amazon, high ratings do not consistently correlate with a high number of reviews. Some books with ratings around 4-5 have a decent number of reviews, but overall engagement remains lower than Amazon's top-rated books.

Contradictions & Possible Explanations

- **Low User Engagement:** Many Jarir customers do not leave reviews.
- **External Ratings:** Users may rely on Goodreads instead of Jarir.
- **Lack of Incentives:** Unlike Amazon, Jarir may not encourage reviews.
- **Platform Differences:** Amazon ranks books by engagement, while Jarir may not.

7. Comparison Between Primary and Secondary Data

7 .1 Overview of Data Sources

Primary Data (Amazon KSA)

- Collected through **web scraping** from Amazon KSA's bestseller book lists.
- Provides **detailed metadata**, including ratings, reviews, and book attributes.
- **Higher user engagement**, as seen from extensive customer reviews.
- Encountered **data extraction challenges**, such as server blocking and incomplete data retrieval.

Secondary Data (Jarir KSA)

- Obtained from **Jarir KSA's online store**.
- Covers a **broader range of books** than Amazon KSA.
- Lacks detailed user interaction data (low review counts).
- Serves as a **complementary dataset** for trend analysis but does not offer deep insights into individual book performance.

7 .2 Key Discrepancies & Alignments

Variable	Discrepancies or Alignments	Comparison & Insights
Price	Alignments\ Low Discrepancy	They align in having a high number of moderately priced books in their bestseller datasets but show a discrepancy as Amazon has a broader price range, while Jarir follows a more structured pricing model. **Possible Explanation:** Jarir's lower price range could be influenced by local pricing strategies, a smaller selection, or currency exchange effects.
Author Influence	High Alignments	The pattern of top authors exists in both datasets. **Possible Explanation:** Bestselling authors tend to dominate across platforms due to their established readership and consistent demand.
Rating	High Discrepancy	Amazon's highly rated books perform better, while Jarir's ratings have less influence. **Possible Explanation:** Amazon users engage more in rating books, while Jarir shoppers may rely more on in-store browsing.
Number of Reviews	High Discrepancy	Amazon's engagement through reviews helps shape bestsellers, while Jarir lacks substantial user feedback. **Possible Explanation:** Lower review counts in Jarir suggest that in-store sales drive popularity more than online engagement.
Genres	High Alignment	Both platforms share similar reader preferences, with "Children & Young Adults," "Self-Help," and "Fiction" as dominant genres. **Possible Explanation:** These genres cater to universal interests, making them popular across platforms.

Variable	Discrepancies or Alignments	Comparison & Insights
Cover Design Influence	Low Discrepancy	Cover attractiveness has a stronger influence on Jarir's audience, but both datasets still contain a significant number of books with unremarkable covers. **Possible Explanation:** In-store browsing at Jarir makes visual appeal more impactful, whereas Amazon's algorithm-driven recommendations rely more on reviews and ratings.
Price and Rating	High Alignment	Both platforms do not show a strong correlation between price and rating, as high-rated books exist across various price points. **Possible Explanation:** Readers tend to rate books based on content quality rather than price, making pricing less influential in rating distribution.
Rating and Number of Reviews	High Alignment	While Amazon shows more reviews for highly rated books, many still have few reviews. Similarly, Jarir has highly rated books across varying review counts, indicating that review volume alone doesn't determine rating strength.

7 .3 Interpretation of Findings

- **Amazon's dataset** is **richer in user engagement** metrics, making it valuable for analyzing customer preferences.
- **Jarir's dataset** provides **broader coverage**, useful for understanding general book availability but lacks the depth needed for detailed consumer behavior insights.
- **Amazon's data suggests a strong online book-buying culture**, while **Jarir's data reflects a more traditional retail model**.
- **The combination of both datasets provides a balanced view** of the Saudi book market, helping us analyze bestseller trends more comprehensively.

Conclusion: While Amazon serves as the **primary data source** for deep insights, Jarir's data acts as a **complementary source**, filling gaps in the broader market perspective.

8. Summary of New Insights and Hypotheses

This section highlights key insights and hypotheses derived from the comparative analysis of Amazon and Jarir datasets. These findings reveal significant patterns in book popularity, pricing, and consumer behavior:

- There is no strong correlation between book ratings and the number of reviews.
- Well-known authors with an established readership have a higher likelihood of their books becoming bestsellers compared to new authors.
- Books with visually appealing covers are more likely to achieve bestseller status.
- Fiction genres, such as Children & Young Adults and General Fiction, are more commonly represented among bestsellers than non-fiction categories like Health, Religion, and Spirituality.
- Mid-range priced books (neither too expensive nor too cheap) are more likely to become bestsellers, as they balance affordability and perceived quality.

9. Data Preprocessing

Objective: to prepare the data for modeling and ensure it is interpretable and suitable for machine learning algorithms.

9 .1 Handling Duplicate Rows and Data Aggregation

To address duplicate rows in the dataset, we first cleaned the Title column by converting the text to lowercase and removing extra spaces. We then identified rows that had the same title and author.

Next, we aggregated these duplicate rows by calculating the average for numeric columns such as Price and Rating, and the sum for Num of Reviews. For text-based columns like Book Type and Genre, we combined the unique values into a single string.

Additionally, a cover image link was included for each processed row, representing the book's cover. After aggregation, the original duplicate rows were removed from the dataset, and the aggregated data was merged back with the cleaned data.

These steps ensured that the dataset became consistent and ready for analysis without redundant duplicates.

9 .2 Normalization of 'Num of Reviews'

We decided to apply normalization to the **Num Of Reviews** column due to the significant variation in its values. To simplify this, we chose the Min-Max scaling method, as it is the most suitable for the task. By normalizing the data to a scale from 0 to 50, we can achieve better differentiation between books with high and low review counts. This ensures a clearer comparison without overly compressing values.To perform the normalization, we used the MinMaxScaler class from the sklearn.preprocessing module, part of the scikit-learn library.

9 .3 Discretization of 'Price'

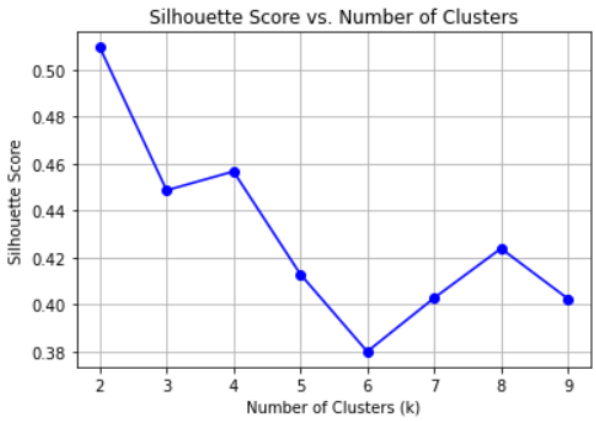
The **Price** column contains continuous values that require simplification. To address this, we performed discretization by dividing the price values into five bins. Each bin corresponds to an ordinal label, ranging from 'Very Low' to 'Very High'. After discretization, we encoded these labels with numeric values from 0 to 4, making them easier to process.

10. Modeling Task :

To explore patterns in book cover designs, we applied clustering techniques using visual features such as **contrast, sharpness, edge complexity,** and **color saturation**.

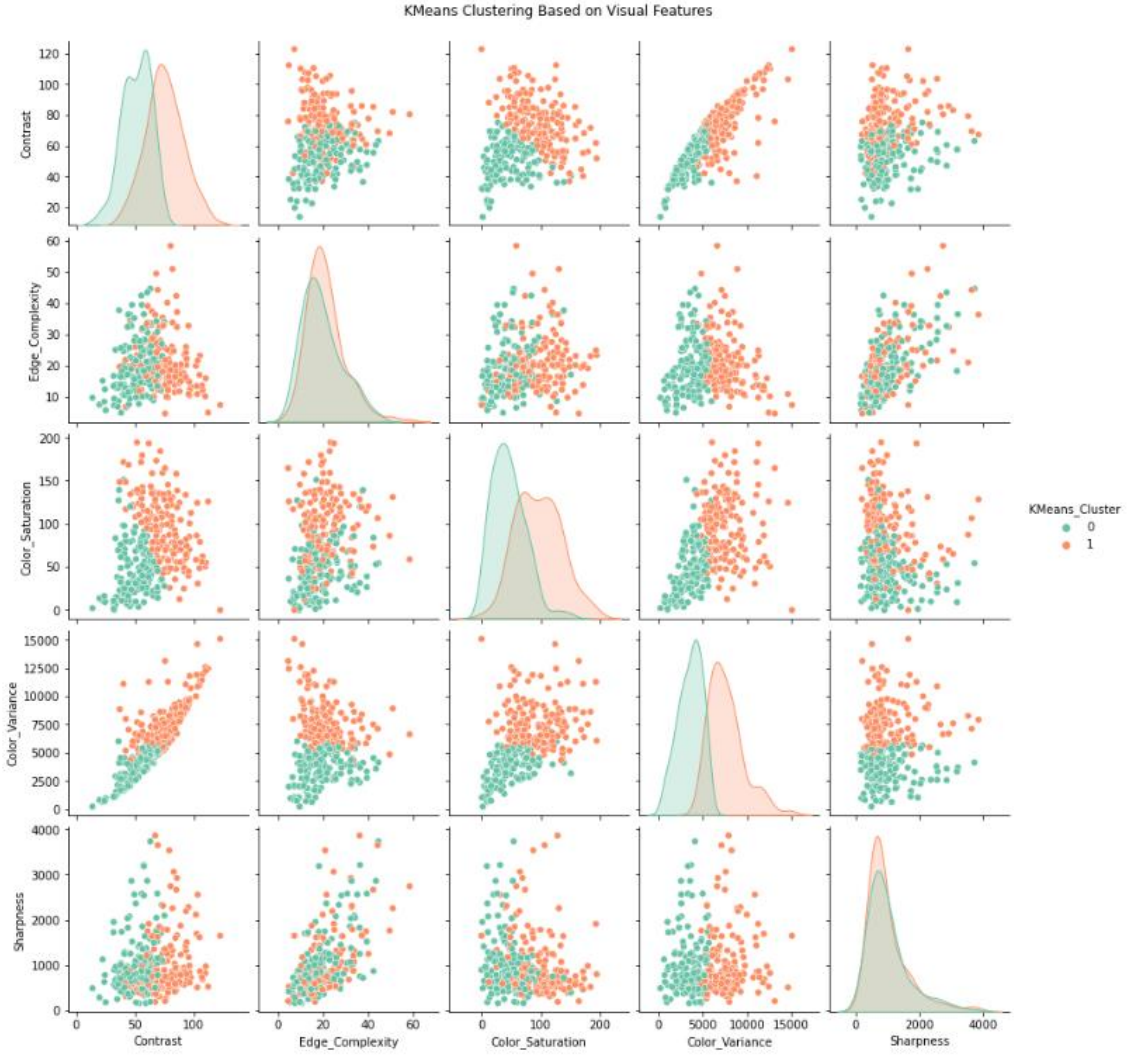
1. Determining the Optimal Number of Clusters:

We used the **Silhouette Score method** as method to find optimal *k* value. The optimal value was found to be **k = 2**, as shown in the figure below.

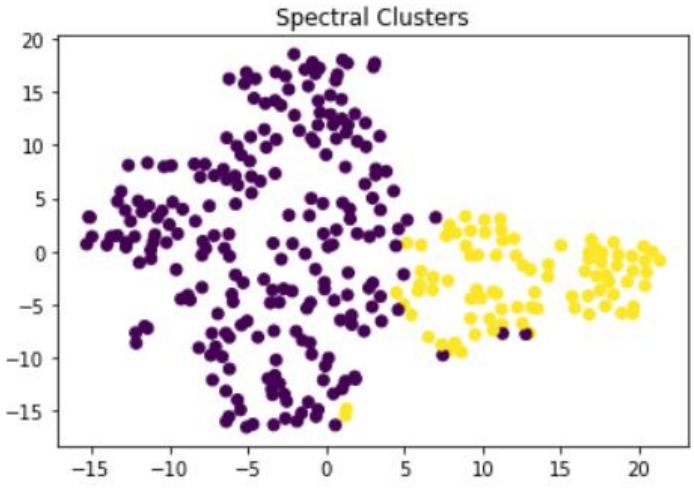


2. Developing Two Clustering Algorithms:

- K-Means Clustering (k = 2)



- **Spectral Clustering (k = 2)**



3. Evaluating the Developed Models:

k-means Cluster Characteristics:

Cluster	Sharpness	Contrast	Color Saturation	Edge Complexity
0	1012.54	51.17	45.55	19.50
1	992.48	75.24	96.30	21.10

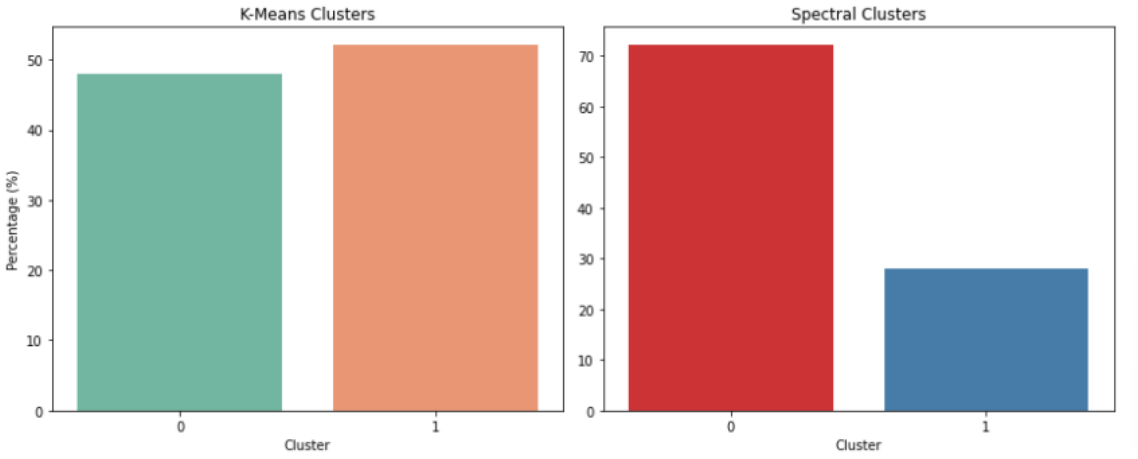
- **Cluster 1:** High contrast, high color saturation, and sharp edges.
- **Cluster 0**: Moderate features, less visually striking.

Spectral Cluster Characteristics:

Cluster	Sharpness	Contrast	Color Saturation	Edge Complexity
0	1094.40	70.45	85.84	22.56
1	763.9	46.30	36.20	14.60

- **Cluster 1:** Lower contrast, muted colors, simpler edges.
- **Cluster 0**: High contrast, high color saturation, and less sharper edges.

k-means spectral clustering distribution:



4. Results summary:

Algorithm	Silhouette Score
K-means	0.25
Spectral Clustering	0.22

Both clustering methods revealed two main groups of cover designs:

- **Cluster 1 (K-Means) / Cluster 0 (Spectral):** Covers with high contrast, vibrant color saturation, and either sharp or soft edges.
- **Cluster 0 (K-Means) / Cluster 1 (Spectral):** Covers with more subtle color use, lower contrast, and simpler designs.

Although the exact grouping of covers differed slightly between methods, both algorithms supported the same general finding:

Bold, high-impact visuals (around 70% of covers) are more dominant among bestsellers, while subtler designs make up the remaining 30%.

Based on the silhouette score and cluster visualization, both models were good but we selected **K-Means** as the better clustering algorithm.

- Works best for spherical, evenly sized clusters.
- Easier to understand since Spectral Clustering uses math tricks (eigenvalues)
- Neither method made great clusters (scores are low), But K-Means was less bad and easier to act on.

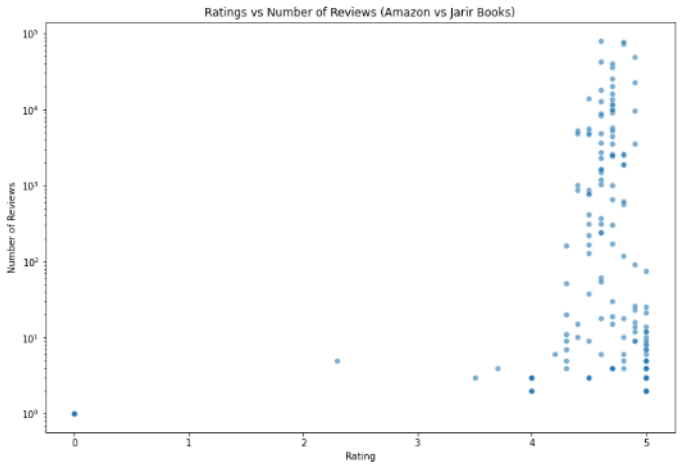
Conclusion of Clustering Task:

This clustering analysis suggests that visual boldness—through contrast and color—is a key pattern among bestsellers. Designers can leverage this insight when crafting new covers.

11. Results and Discussion:

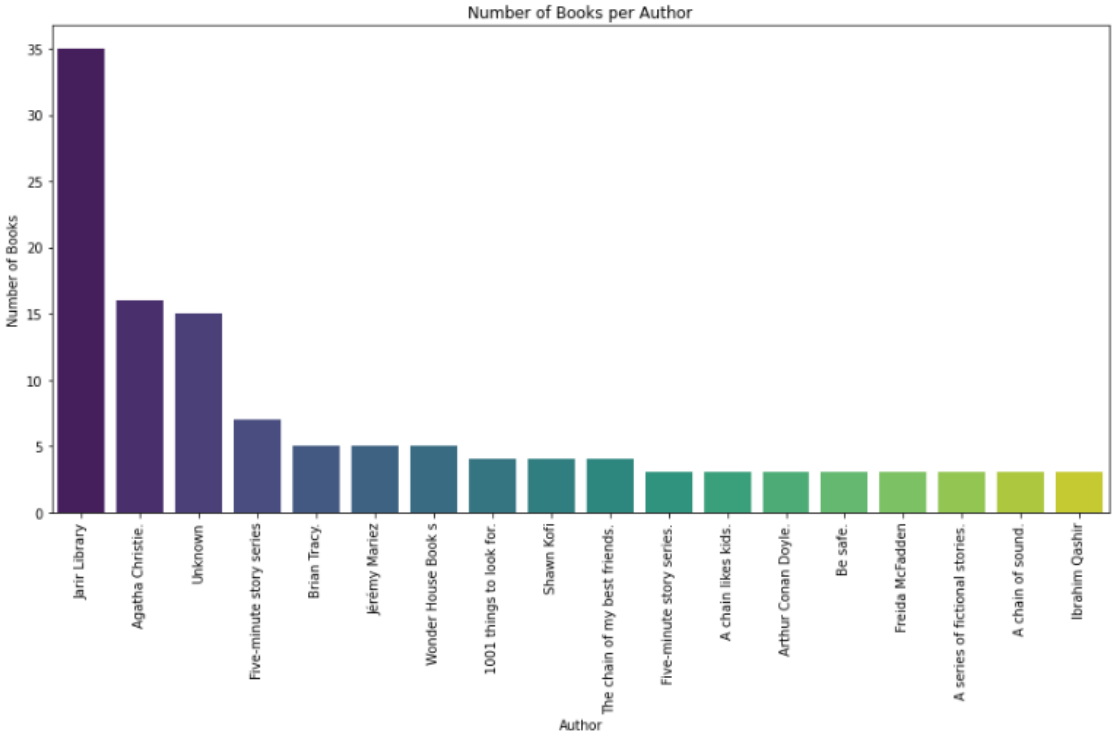
The overall bestseller books characteristics that we aim to study from the beginning, to contribute valuable insights to the publishing industry and help stakeholders make data-driven decisions are as follow:

1. How do ratings and the number of reviews vary among bestsellers?



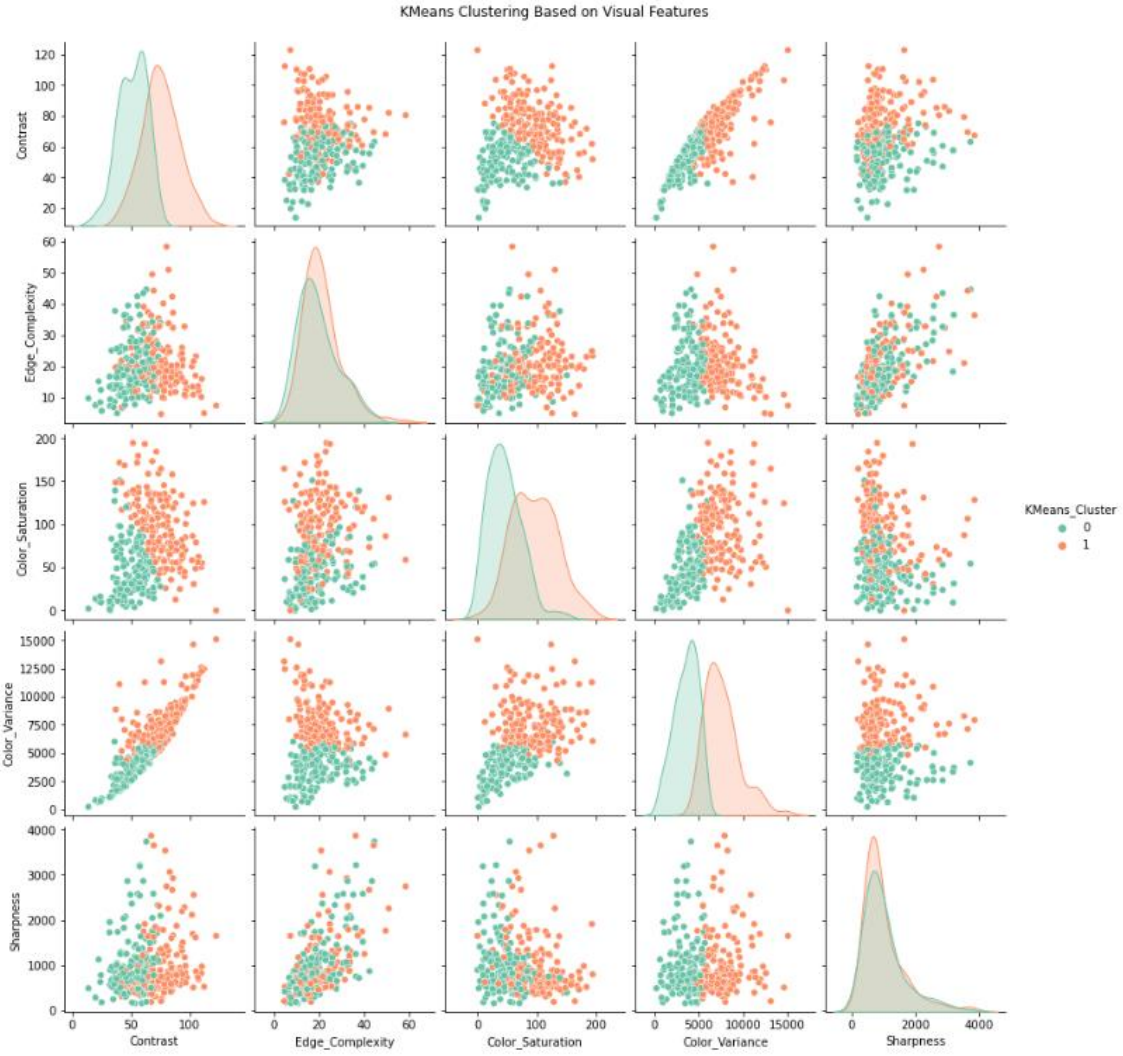
- **No strong** correlation between book ratings and the number of reviews.
- The spread of review counts is **inconsistent**, suggesting that high ratings do not necessarily translate into more reviews.
- Lower-rated books (below 4 stars) receive almost no reviews, reinforcing the trend that books with more visibility and popularity tend to accumulate higher engagement, while **lesser-known ones struggle for attention.**

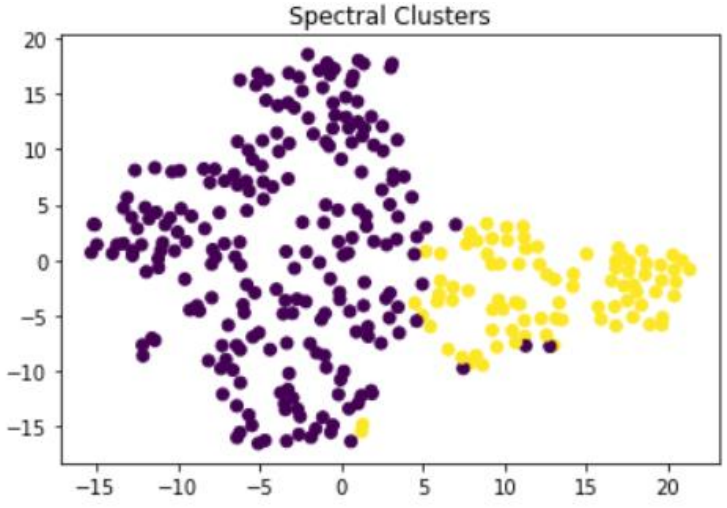
2. Are certain authors more likely to have their books become bestsellers?



- The graph highlights that bestsellers are **significantly influenced by authors**, as many of them have multiple books appearing in the dataset. This trend suggests that well-known authors often sustain a strong presence in the bestseller list due to their established reputation and consistent readership, which is consistent with the multivariate analysis showing that authors with multiple bestsellers tend to maintain strong sales across their books such as: Jarir and Agatha Christie.

3. What visual design patterns are consistently associated with bestseller book covers?





We applied unsupervised clustering using two algorithms—K-Means and Spectral Clustering—on image-based features (contrast, color saturation, sharpness, etc.) of bestseller book covers. The goal was to discover natural groupings in cover design styles and understand which patterns are most common.

Both clustering methods revealed two main groups of cover designs:

- **Cluster 1 (K-Means) / Cluster 0 (Spectral):** Covers with high contrast, vibrant color saturation, and either sharp or soft edges.
- **Cluster 0 (K-Means) / Cluster 1 (Spectral):** Covers with more subtle color use, lower contrast, and simpler designs.

Although the exact grouping of covers differed slightly between methods, both algorithms supported the same general finding:

Bold, high-impact visuals (around 70% of covers) are more dominant among bestsellers, while subtler designs make up the remaining 30%.

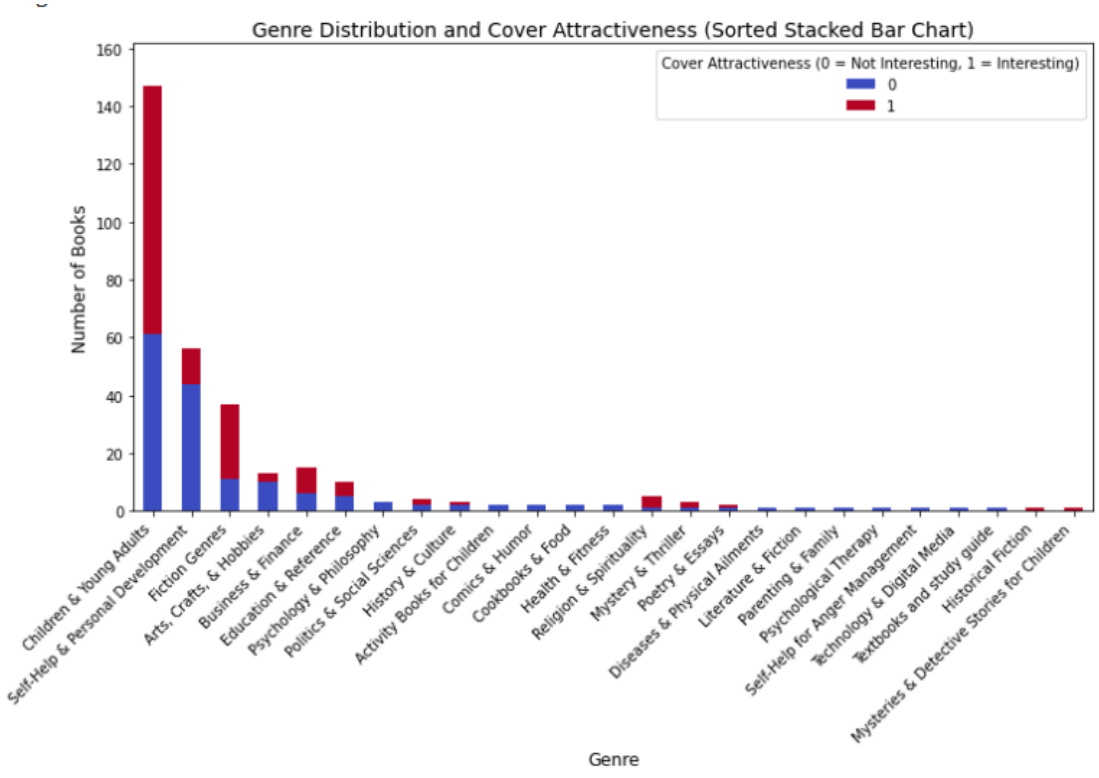
Algorithm	Silhouette Score
K-means	0.25
Spectral Clustering	0.22

Based on the silhouette score and cluster visualization, both models were good but we selected **K-Means** as the better clustering algorithm.

- Works best for spherical, evenly sized clusters.
- Easier to understand since Spectral Clustering uses math tricks (eigenvalues)
- Neither method made great clusters (scores are low), But K-Means was less bad and easier to act on.

Finally, this clustering analysis suggests that visual boldness—through contrast and color—is a key pattern among bestsellers. Designers can leverage this insight when crafting new covers.

4. What genres are most represented among bestsellers?



The analysis reveals a clear relationship between genre and bestseller status genres like *Children & Young Adults*, *Self-Help*, and *Fiction* consistently dominate in volume, suggesting that certain genres inherently have broader appeal and higher market demand, making them more likely to produce bestsellers.

5. What is the relationship between price and bestseller books? What is the price range of bestseller books?

- **Likely Negative Correlation:** Bestseller books appear to cluster in the lower-to-mid price ranges (20 – 100), as this range typically maximizes accessibility and sales volume.
- **Limited High-Priced Bestsellers:** Fewer bestsellers are observed at the upper price range (more than 150), suggesting premium pricing may reduce mass-market appeal.
- The concentration of dots/strip plots at lower prices supports the idea that affordability drives bestseller status.

12. Conclusion and Future Work:

This study provides a comprehensive, data-driven exploration of the key factors associated with bestseller books, offering actionable insights for authors, publishers, and designers. By examining ratings, authorship, visual design, genre trends, and pricing, we uncovered meaningful patterns that contribute to a book's success in the market.

Our findings reveal that bestseller status is influenced by a combination of factors rather than a single element. While no strong correlation exists between ratings and review counts, popular authors often dominate the bestseller space due to loyal readership. Visually, bold and vibrant cover designs—characterized by strong contrast, sharpness, and color saturation—are more prevalent among bestsellers, reinforcing the importance of eye-catching visuals in marketing.

Genre also plays a critical role, with categories like *Children & Young Adults*, *Self-Help*, and *Fiction* emerging as consistently dominant. In terms of pricing, books in the lower-to-mid price range (20–100) tend to perform better, likely due to greater affordability and accessibility for a wider audience.

Together, these insights highlight that bestseller success is shaped by strategic decisions across multiple dimensions—authorship, design, genre targeting, and pricing. Stakeholders in the publishing industry can use these patterns to refine their content strategies, optimize visual presentation, and align offerings with market demand.

Ultimately, this research supports the value of data analytics in understanding consumer = behavior and enhancing the visibility and impact of published works.

Future Work

While this study offers valuable insights, several opportunities exist for further exploration. Future research could:

- **Expand data sources** to include international markets, independent publishers, or other retail platforms for broader generalization.

- **Incorporate sentiment analysis** from user reviews to better understand reader preferences and emotional reactions tied to bestseller performance.
- **Apply deep learning models** (e.g., convolutional neural networks) to extract more complex visual features from book covers for improved clustering accuracy.
- **Investigate temporal patterns**, such as seasonal trends or publication timing, to assess how timing influences bestseller status.
- **Explore cross-platform performance**, comparing how the same books perform across different vendors (e.g., Amazon, Jarir, Goodreads).
- **Integrate social media trends** or influencer mentions as predictive variables to understand the impact of online visibility on sales.

These directions can further enhance the predictive power and strategic value of data analytics in the publishing industry.

13. Challenges:

Data collection comes with various challenges that can hinder efficiency and accuracy. In our process, which involves web scraping, we faced several key difficulties:

1. **Time-Consuming Process:**
   Data collection, especially when using web scraping techniques, requires significant time due to the complexity of extracting and processing data from multiple sources.
2. **Unclear HTML Structure:**
   Some essential elements like and do not have clear or consistent class names, making it difficult to identify and extract the required data efficiently.
3. **Dynamic Content with JavaScript:**
   Certain websites load content dynamically using JavaScript, which means that the data may not be visible in the initial HTML source code. This requires additional tools or techniques to handle dynamic content effectively.
4. **Request Limits and Access Restrictions:**
   Some data sources impose strict limits on the number of requests that can be made

within a specific timeframe, while others require special access permissions or API keys.

5. **Inconsistent Data Availability:**
   Some information is available in certain sources but missing in others, leading to incomplete datasets and making it challenging to ensure data consistency and reliability.
6. **Choosing the Best Plot for Multivariate Data:**
   Selecting the most appropriate visualization for multivariate data was challenging, as different types of variables required different graphical representations.
7. **Difficulty in Identifying a Suitable Model for Classifying Book Cover Interest.**
8. **Variability in Book Cover Image Quality:**
   Some book cover images were of low resolution or highly compressed, which impacted the accuracy of image-based feature extraction.
9. **Choose modeling task:**
   Struggled to identify a suitable modeling task and algorithm that would yield meaningful insights.

14. Recommendations for Mitigating Challenges:

To address the challenges encountered during data collection, particularly in web scraping, the following recommendations can help improve efficiency and accuracy:

1. **Optimize the Scraping Process:**
   - Use efficient web scraping tools and libraries like **BeautifulSoup** with asynchronous requests to enhance speed.
2. **Handling Unclear HTML Elements:**
   - Focus on attributes like **ID** and the general structure of elements rather than relying solely on class names.
   - Use web page analysis tools to identify required elements for data extraction efficiently.

3. **Dealing with Dynamic Content:**
    - Utilize tools that support dynamic page loading, such as **Selenium**, to handle content that loads asynchronously.
4. **Implement Time Delays Between Requests:**
    - Introduce time delays between requests to mimic natural user behavior, reducing the chances of being blocked by the server.
5. **Utilize Ready-Made Data Scraping Tools:**
    - Use tools like **Instant Data Scraper** to quickly obtain an initial dataset before refining the extraction process.
    - Instead of sending excessive requests to a single page (which may trigger request limits), gather links to individual pages, organize them in a spreadsheet, and access them separately. This method reduces server load and minimizes the risk of hitting request limits.
6. **Addressing Incomplete Data:**
    - **Gather Data from Multiple Sources:** Combine information from different sources to fill in missing data and ensure completeness.
    - **Manual Data Review and Completion:** Conduct manual reviews and validation processes to identify and complete missing information when automated methods fall short.
7. **Search recommendations** on suitable graphs based on variable types.
8. **Conducted a literature review** and explored different machine learning techniques: Since no prior studies classified book cover interest, we opted for unsupervised learning to uncover patterns in the dataset.
9. **Applied denoising techniques**:
   Including sharpening filters such as the Laplacian filter, to enhance image quality before analysis.
10. Explored **multiple modeling approaches** and experimented with various algorithms to determine the most effective solution.

15. References

1. A. Alharbi, "Exploring Factors Influencing the Amazon Best-Selling Books Selection Process from 2009 to 2019," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/382998978_Exploring_Factors_Influencing_the_Amazon_Best-Selling_Books_Selection_Process_from_2009_to_2019

2. J. Smith and J. Doe, "Using Full-Text Content to Characterize and Identify Best Seller Books," PLOS ONE, May 11, 2023. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0302070

3. L. Johnson and K. Brown, "Analyzing Social Book Reading Behavior on Goodreads and How It Predicts Amazon Best Sellers," ResearchGate, 2018. [Online]. Available: https://www.researchgate.net/publication/327789907_Analyzing_Social_Book_Reading_Behavior_on_Goodreads_and_how_it_predicts_Amazon_Best_Sellers.

4. "Amazon Book – 100 Best Sellers," Amazon Saudi Arabia. [Online]. Available: https://www.amazon.sa/-/en/gp/bestsellers/books/ref=zg_bs_pg_2_books?ie=UTF8&pg=2 . [Accessed: Jan. 31, 2025].

5. "Jarir Book – Arabic Books Best Sellers," Jarir Bookstore. [Online]. Available: https://www.jarir.com/sa-en/arabic-books.html?producttag=Best%20Seller#listingContent. [Accessed: Feb. 1, 2025].

6. "Jarir Book – English Books Best Sellers," Jarir Bookstore. [Online]. Available: https://www.jarir.com/sa-en/english-books.html?producttag=Best%20Seller#listingContent. [Accessed: Feb. 2, 2025].

7. "Web Scraper - Free Web Scraping," Google Chrome Web Store. [Online]. Available: https://chromewebstore.google.com/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn?hl=en

8. "Instant Data Scraper - Free Web Scraping," Google Chrome Web Store. [Online]. Available: https://chromewebstore.google.com/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah

9. Hurix Digital, "Knowing the Psychology of Book Cover Design," Hurix Blogs, 2023. [Online]. Available: https://www.hurix.com/blogs/knowing-the-psychology-of-book-cover-design