Book Review Analysis Report

---

1. Introduction

In the era of data-driven decision-making, bookstores must leverage insights from customer preferences to optimize inventory, marketing, and sales strategies. This project analyzes the **Book-Crossing dataset**, a comprehensive collection of book ratings and user demographics, to uncover valuable trends and recommendations for a bookstore business. By identifying high-demand books, regional preferences, and hidden gems, this analysis provides actionable insights to enhance customer engagement and drive sales.

---

2. Dataset Overview

The dataset, sourced from Kaggle, includes:

- **278,858 users** (anonymized, with demographic data such as age, location, and language).
- **1,149,780 ratings** (both explicit and implicit) for books.
- **271,379 books** across various genres, authors, and publishers.
- Structured into three normalized tables: **users**, **books**, and **ratings**, making it well-suited for SQL-based analysis.

---

3. Business Case

The analysis addresses the following key business objectives:

1. **Identify high-demand books** by analyzing rating trends across genres, languages, and publication years.
2. **Optimize inventory** for different store locations based on regional reading preferences.
3. **Analyze customer demographics** to recommend books tailored to age groups and interests.
4. **Discover hidden gems**—books with high ratings but low visibility—that could become niche bestsellers with targeted promotion.

---

4. Analytical Approach

To address these objectives, the following analytical questions were formulated:

1. What are the most popular books based on ratings?
2. Which genres or categories have the highest ratings and engagement?
3. What are the reading preferences in different locations?
4. Do younger and older readers prefer different genres?
5. Are there books with high ratings but low visibility (hidden gems)?
6. How do book preferences change over time? Are newer books rated better than older ones?
7. Which publishers are the most popular based on ratings?

---

5. Methodology

The analysis was conducted using **SQL** for data extraction and **Power BI** for visualization. Key steps included:

- **Data Cleaning**: Rigorous preprocessing to handle missing values, standardize text, and ensure data integrity.
- **Weighted Rating Calculation**: Adapted IMDb's Bayesian average formula to rank books fairly, balancing average ratings and review counts.
- **Geographical Analysis**: Identified regional preferences by analyzing user locations and book ratings.
- **Demographic Insights**: Segmented users by age groups to uncover genre preferences.

---

6. Data Cleaning

Before performing analysis, a rigorous data cleaning process was conducted to ensure accuracy and consistency.

6.1 Database and Table Setup

The dataset consists of three main tables:

- **Users**: Contains user demographics, including location and age.
- **Books**: Holds book details such as title, author, and publisher.

- **Ratings**: Stores user-provided book ratings.

```sql
CREATE TABLE users (
    user_id INT PRIMARY KEY,
    location TEXT,
    age INT
);
```

```sql
CREATE TABLE books (
    isbn VARCHAR(20) PRIMARY KEY,
    book_title TEXT,
    book_author TEXT,
    year_of_publication INT,
    publisher TEXT,
    img_s TEXT,
    img_m TEXT,
    img_l TEXT
);
```

```sql
CREATE TABLE ratings (
    user_id INT,
    isbn VARCHAR(30),
    book_rating INT CHECK (book_rating BETWEEN 0 AND 10),
    PRIMARY KEY (user_id, isbn),
    FOREIGN KEY (user_id) REFERENCES users(user_id) ON DELETE CASCADE,
    FOREIGN KEY (isbn) REFERENCES books(isbn) ON DELETE CASCADE
```

```
);
```

6.2 Handling Missing and Incorrect Data

- **Ratings**: Ratings of 0 were treated as missing values and replaced with NULL.
- **Language and Category**: Missing values encoded as '9' were replaced with NULL.
- **ISBN Formatting**: Fixed formatting issues by removing slashes and extra spaces.
- **Standardizing Text Data**: Book titles and author names were converted to lowercase for consistency.

```sql
UPDATE ratings SET book_rating = NULL WHERE book_rating = 0;

UPDATE books SET language = NULL WHERE language = '9';

UPDATE books SET category = NULL WHERE category = '9';

UPDATE ratings SET isbn = TRIM(REPLACE(isbn, '/', '')) WHERE isbn LIKE '%/%';

UPDATE books SET isbn = TRIM(REPLACE(isbn, '/', '')) WHERE isbn LIKE '%/%';

UPDATE books SET book_title = LOWER(book_title), book_author = LOWER(book_author);
```

6.3 Cleaning the Users Table

- **Location Column**: Split into City, State, and Country for improved usability.
- **Standardizing Country Names**: A mapping table was created to standardize country names.

```sql
ALTER TABLE users

ADD COLUMN city VARCHAR(255),

ADD COLUMN state VARCHAR(255),

ADD COLUMN country VARCHAR(255);
```

```sql
UPDATE users SET city = SPLIT_PART(location, ',', 1),
                 state = SPLIT_PART(location, ',', 2),
                 country = SPLIT_PART(location, ',', 3);
```

```sql
CREATE TABLE country_mapping (
    raw_country VARCHAR(255) PRIMARY KEY,
    standard_country VARCHAR(255));
```

```sql
UPDATE users
SET cleaned_country = COALESCE(
    m.standard_country,
    TRIM(LOWER(users.country)),
    'Unknown'
)
FROM country_mapping m
WHERE LOWER(TRIM(users.cleaned_country)) =
LOWER(TRIM(m.raw_country))
   OR LOWER(TRIM(users.country)) = LOWER(TRIM(m.raw_country));
```

---

7. Analysis and Results

7.1 Most Popular Books

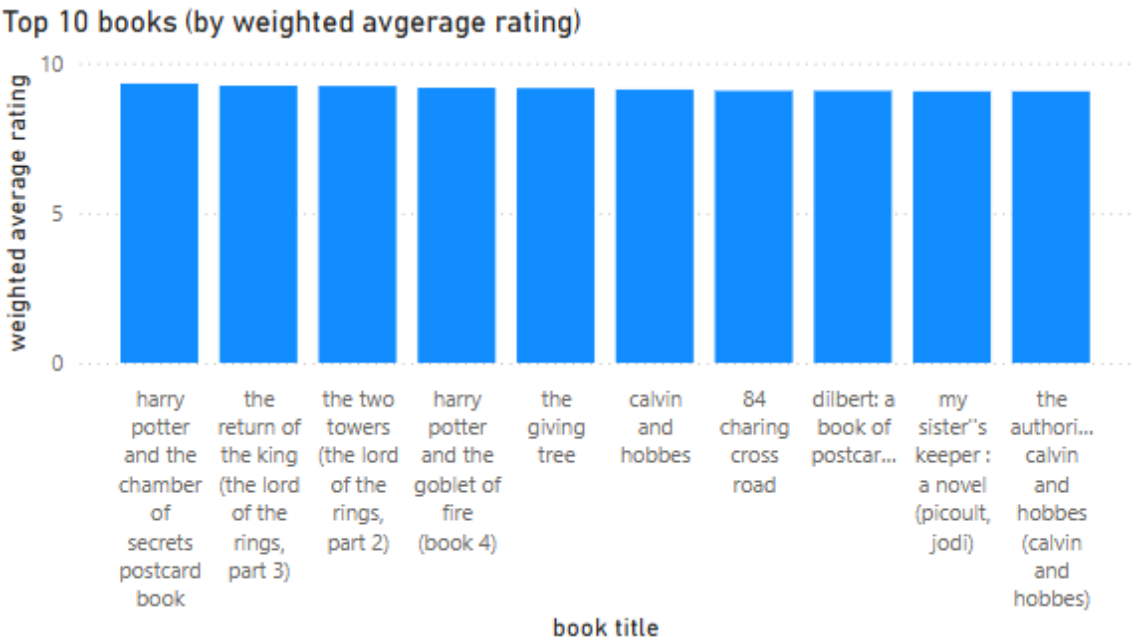
To identify the most popular books, we used a weighted rating formula to balance average ratings and the number of reviews.

```sql
WITH rating_stats AS (
    SELECT
```

```
        ISBN,
        COUNT(book_rating) AS num_ratings,
        AVG(book_rating) AS avg_rating
    FROM ratings
    WHERE book_rating > 0
    GROUP BY ISBN
), global_avg AS (
    SELECT
        AVG(avg_rating)::NUMERIC AS C,
        PERCENTILE_CONT(0.95) WITHIN GROUP (ORDER BY
num_ratings)::NUMERIC AS m
    FROM rating_stats
)
SELECT
    b.ISBN,
    b.book_title,
    b.book_author,
    rs.num_ratings,
    ROUND(rs.avg_rating::NUMERIC, 2) AS avg_rating,
    ROUND(((rs.num_ratings / (rs.num_ratings + g.m) *
rs.avg_rating) +
          (g.m / (rs.num_ratings + g.m) * g.C))::NUMERIC, 2)
AS weighted_score
FROM rating_stats rs
JOIN books b ON rs.ISBN = b.ISBN
CROSS JOIN global_avg g
ORDER BY weighted_score DESC
LIMIT 10;
```

Results (Partial View of the Data):

ISBN	Book Title	Book Author	Num Ratings	Avg Rating	Weighted Score
0439425220	Harry Potter and the Chamber of Secrets Postcard Book	J.K. Rowling	23	9.87	9.32
0345339738	The Return of the King (The Lord of the Rings, Part 3)	J.R.R. Tolkien	77	9.40	9.25
0618002235	The Two Towers (The Lord of the Rings, Part 2)	J.R.R. Tolkien	25	9.72	9.24



Conclusion: The top-rated books include **Harry Potter** and **The Lord of the Rings** series, reflecting their widespread popularity and high reader engagement.

7.2 Genre Preferences

To analyze genre preferences, we calculated weighted scores for each category.

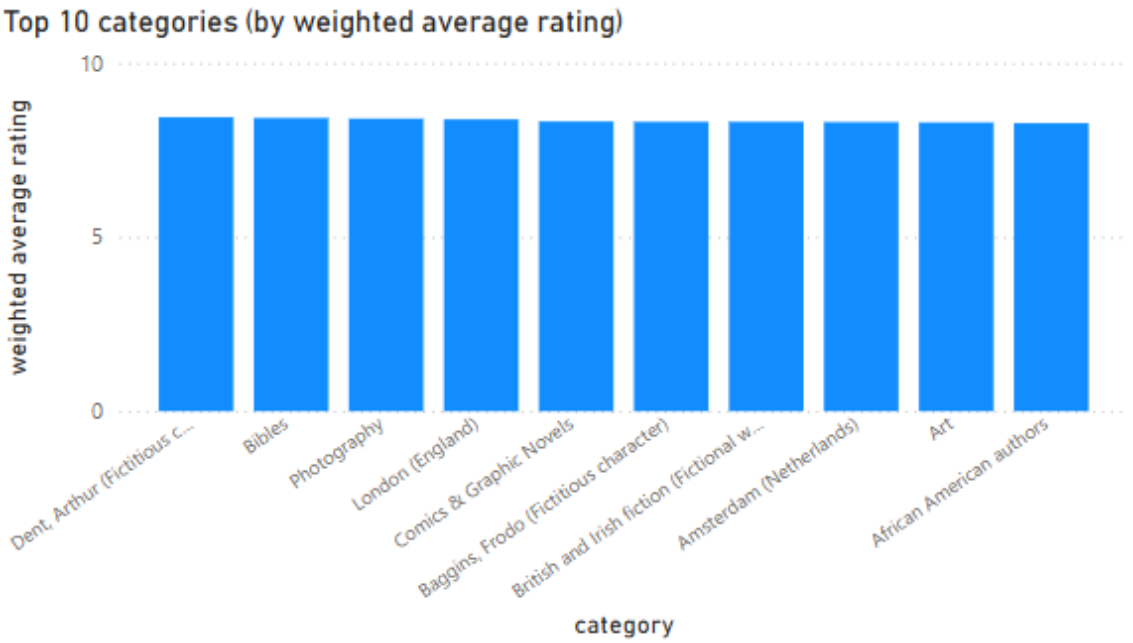
```sql
WITH category_stats AS (

    SELECT
```

```
        b.category,
        COUNT(r.book_rating) AS num_ratings,
        AVG(r.book_rating) AS avg_rating
    FROM ratings r
    JOIN books b ON r.ISBN = b.ISBN
    WHERE r.book_rating > 0
    GROUP BY b.category
), global_avg AS (
    SELECT
        AVG(avg_rating)::NUMERIC AS C,
        PERCENTILE_CONT(0.95) WITHIN GROUP (ORDER BY
num_ratings)::NUMERIC AS m
    FROM category_stats
)
SELECT
    cs.category,
    cs.num_ratings,
    ROUND(cs.avg_rating, 2) AS avg_rating,
    ROUND(((cs.num_ratings / (cs.num_ratings + g.m) *
cs.avg_rating) +
          (g.m / (cs.num_ratings + g.m) * g.C))::NUMERIC, 2)
AS weighted_score
FROM category_stats cs
CROSS JOIN global_avg g
ORDER BY weighted_score DESC
LIMIT 10;
```

Results (Partial View of the Data):

Category	Num Ratings	Avg Rating	Weighted Score
Dent, Arthur (Fictitious character)	91	8.75	8.44
Bibles	34	9.24	8.42
Photography	141	8.59	8.40



Conclusion: Multiple categories, including **science fiction (Arthur Dent)**, **Bibles**, **Photography**, and **Comics & Graphic Novels**, have high weighted average ratings. This suggests strong reader appreciation across diverse genres.

7.3 Regional Preferences

To identify regional reading preferences, we analyzed the most-rated books by country.

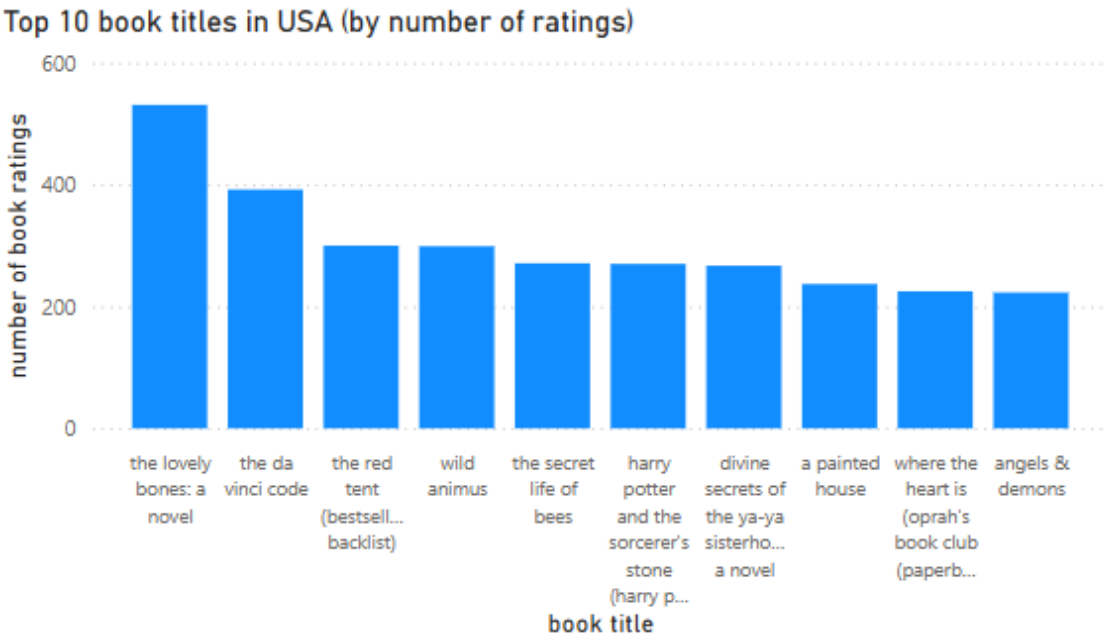
```sql
WITH TopCountries AS (
    SELECT
        country,
        COUNT(user_id) AS user_count
    FROM users
    WHERE country IS NOT NULL
```

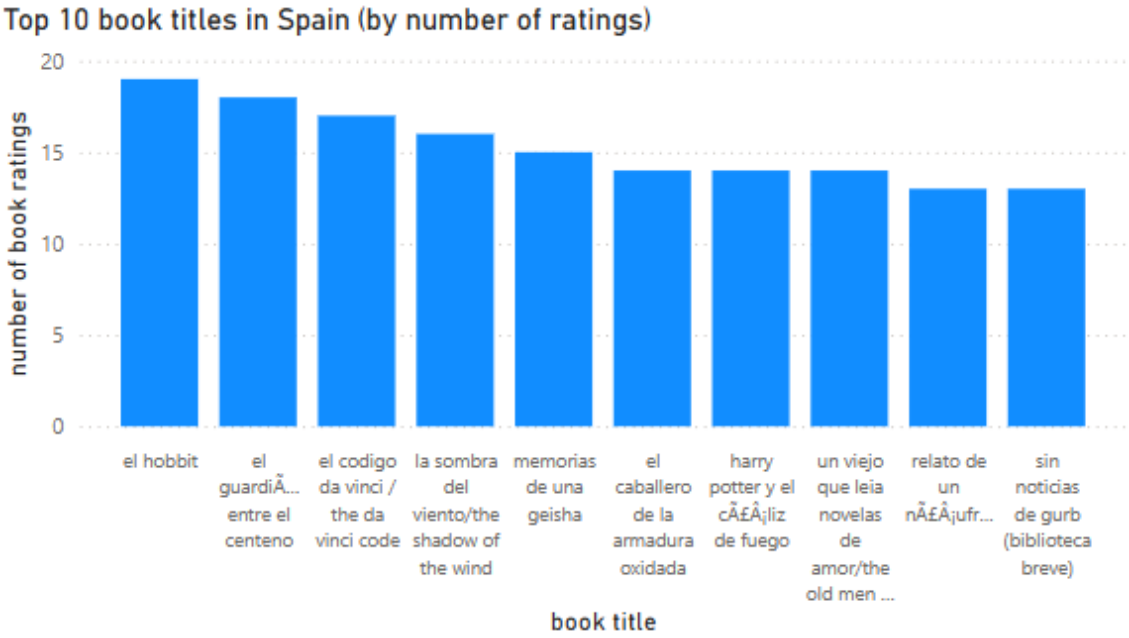
```
    GROUP BY country
    ORDER BY user_count DESC
    LIMIT 10
),
CountryRatings AS (
    SELECT
        u.country,
        r.ISBN,
        b.book_title,
        COUNT(r.book_rating) AS rating_count
    FROM ratings r
    JOIN users u ON r.user_id = u.user_id
    JOIN books b ON r.ISBN = b.ISBN
    WHERE u.country IN (SELECT country FROM TopCountries)
    GROUP BY u.country, r.ISBN, b.book_title
),
RankedBooks AS (
    SELECT
        country,
        ISBN,
        book_title,
        rating_count,
        RANK() OVER (PARTITION BY country ORDER BY
rating_count DESC) AS rank
    FROM CountryRatings
)
SELECT
    country,
    ISBN,
    book_title,
    rating_count
FROM RankedBooks
```

```
WHERE rank <= 5
ORDER BY country, rank;
```

Results(Partial View of the Data):

Country	ISBN	Book Title	Rating Count
USA	0316666343	The Lovely Bones: A Novel	531
USA	0385504209	The Da Vinci Code	392
Canada	0316666343	The Lovely Bones: A Novel	62





Conclusion: **The Lovely Bones** and **The Da Vinci Code** are highly popular in the USA and Canada, reflecting regional preferences.

7.4 Age Group Preferences

To analyze genre preferences by age group, we segmented users into "Under 18", "18-30", "31-50", and "51+".

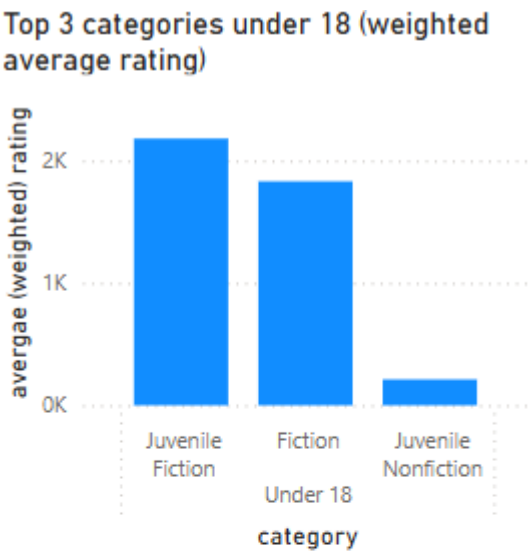
```sql
WITH user_age_groups AS (
    SELECT
```

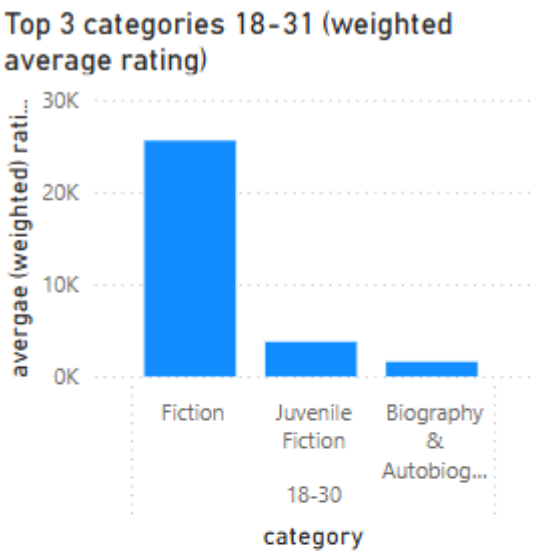
```
        user_id,
        CASE
            WHEN age IS NULL THEN 'Unknown'
            WHEN age < 18 THEN 'Under 18'
            WHEN age BETWEEN 18 AND 30 THEN '18-30'
            WHEN age BETWEEN 31 AND 50 THEN '31-50'
            ELSE '51+'
        END AS age_group
    FROM users
), category_stats AS (
    SELECT
        u.age_group,
        b.category,
        COUNT(r.book_rating) AS num_ratings
    FROM ratings r
    JOIN user_age_groups u ON r.user_id = u.user_id
    JOIN books b ON r.ISBN = b.ISBN
    WHERE r.book_rating > 0
    GROUP BY u.age_group, b.category
), ranked_categories AS (
    SELECT
        age_group,
        category,
        num_ratings,
        RANK() OVER (PARTITION BY age_group ORDER BY
num_ratings DESC) AS rank
    FROM category_stats
     WHERE category is not null
)
SELECT age_group, category, num_ratings
FROM ranked_categories
WHERE rank <= 5
```

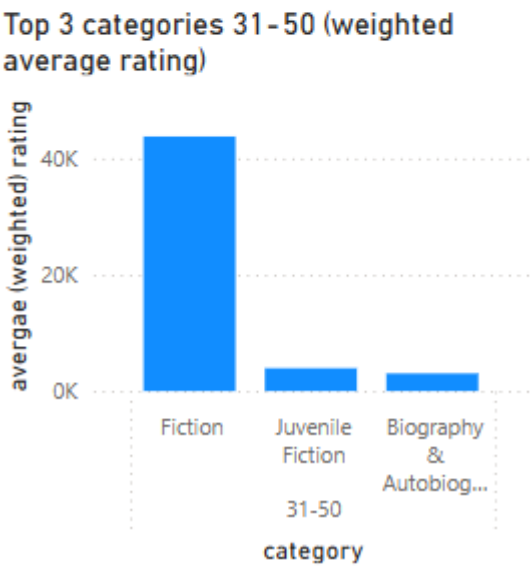
```
ORDER BY age_group, rank;
```

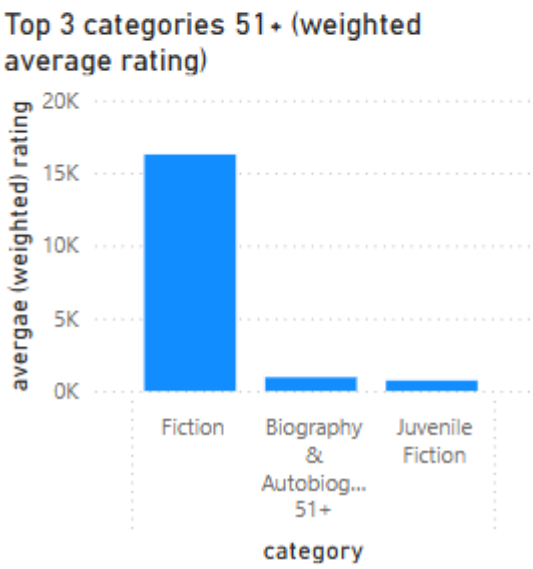
Results(Partial View of the Data):

Age Group	Category	Num Ratings
18-30	Fiction	25464
18-30	Juvenile Fiction	3735
31-50	Fiction	43604
31-50	Juvenile Fiction	3894









Conclusion: **Fiction** is the most popular category across all age groups, with **Juvenile Fiction** also being highly rated among younger readers.

7.5 Hidden Gems

To identify hidden gems, we filtered books with high ratings but low visibility.

```sql
WITH BookRatingCounts AS (
    SELECT
        b.book_title,
        COUNT(r.book_rating) AS rating_count,
        AVG(r.book_rating) AS avg_rating
    FROM ratings r
     JOIN books b ON b.isbn = r.isbn
    GROUP BY book_title
)
SELECT
    book_title,
    avg_rating,
    rating_count
FROM BookRatingCounts
WHERE
    rating_count BETWEEN 5 AND 50
    AND avg_rating = 10
ORDER BY avg_rating DESC, rating_count ASC;
```

Results:

Book Title	Avg Rating	Rating Count
A Voice in the Wind (Mark of the Lion #1)	10.00	5
The Little Zen Companion	10.00	5

Bury me with soldiers: one grunt's honest story about vietnam	10.00	5

Conclusion: These books have high ratings but low visibility, making them ideal candidates for targeted promotion.

7.6 Book Preferences Over Time

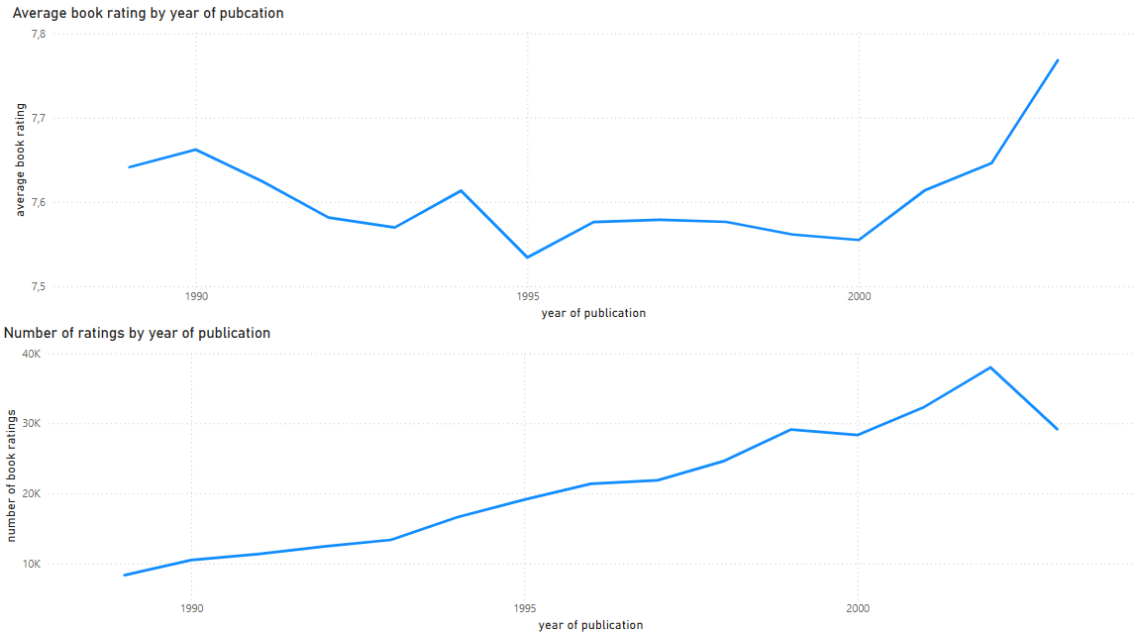
To analyze how book preferences change over time, we grouped ratings by publication year.

```sql
WITH book_stats AS (
    SELECT
        b.year_of_publication,
        COUNT(r.book_rating) AS num_ratings,
        AVG(r.book_rating) AS avg_rating
    FROM ratings r
    JOIN books b ON r.isbn = b.isbn
    WHERE r.book_rating > 0 AND b.year_of_publication IS NOT NULL
    GROUP BY b.year_of_publication
), min_ratings_threshold AS (
    SELECT
        PERCENTILE_CONT(0.85) WITHIN GROUP (ORDER BY num_ratings) AS min_reviews
    FROM book_stats
)
SELECT
    bs.year_of_publication,
    bs.num_ratings,
    ROUND(bs.avg_rating, 2) AS avg_rating
FROM book_stats bs
JOIN min_ratings_threshold mrt ON bs.num_ratings >= mrt.min_reviews
```

```
ORDER BY bs.year_of_publication;
```

Results (Partial View of the Data):

Year of Publication	Num Ratings	Avg Rating
2000	28321	7.55
32331	32331	7.61
2002	37986	7.65



Conclusion: There is an increasing number of ratings over time and a stable average rating with a slight increase in later years.

7.7 Publisher Performance

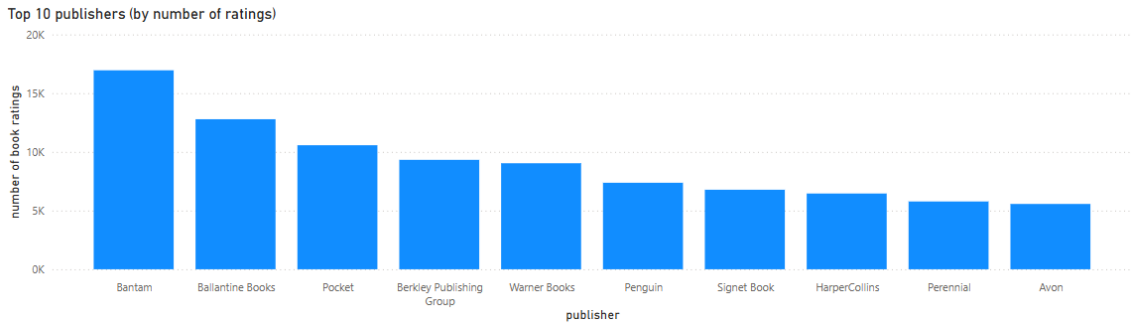
To identify the most popular publishers, we calculated weighted scores based on ratings.

```sql
WITH publisher_stats AS (
    SELECT
        b.publisher,
```

```
        COUNT(r.book_rating) AS num_ratings,
        AVG(r.book_rating) AS avg_rating
    FROM ratings r
    JOIN books b ON r.isbn = b.isbn
    WHERE r.book_rating > 0 AND b.publisher IS NOT NULL
    GROUP BY b.publisher
), global_avg AS (
    SELECT
        AVG(avg_rating)::NUMERIC AS C,
        PERCENTILE_CONT(0.95) WITHIN GROUP (ORDER BY
num_ratings)::NUMERIC AS m
    FROM publisher_stats
)
SELECT
    ps.publisher,
    ps.num_ratings,
    ROUND(ps.avg_rating, 2) AS avg_rating,
    ROUND(((ps.num_ratings / (ps.num_ratings + g.m) *
ps.avg_rating) +
          (g.m / (ps.num_ratings + g.m) * g.C))::NUMERIC, 2)
AS weighted_score
FROM publisher_stats ps
CROSS JOIN global_avg g
ORDER BY weighted_score DESC
LIMIT 10;
```

Results:

Publisher	Num Ratings	Avg Rating	Weighted Score
Arthur A. Levine Books	338	8.91	8.73
TokyoPop	375	8.78	8.63
Five Star (ME)	185	8.82	8.55



Conclusion: **Arthur A. Levine Books** and **TokyoPop** are the highest-rated publishers, indicating strong reader satisfaction.

---

8. Conclusion and Business Recommendations

This analysis of the Book-Crossing dataset provides key insights into customer reading preferences, enabling data-driven decision-making for inventory management, marketing strategies, and customer engagement. The key takeaways include:

- **Popular Books and Genres: Bestselling books such as *Harry Potter* and *The Lord of the Rings* continue to dominate reader interest, while Fiction remains the most preferred genre across all age groups. Investing in these categories and ensuring their availability across store locations can maximize sales.**
- **Regional Preferences: Preferences vary significantly by country, with books like *The Lovely Bones* and *The Da Vinci Code* performing exceptionally well in North America. Tailoring inventory and marketing strategies to regional preferences can enhance customer satisfaction and increase store profitability.**
- **Demographic Insights: Younger readers (under 30) show strong engagement with Juvenile Fiction, while older audiences prefer classic and literary fiction. Personalized book recommendations based on age demographics can improve customer retention.**
- **Hidden Gems Strategy: High-rated but lesser-known books present an opportunity for bookstores to introduce curated collections, increasing demand through targeted promotions and in-store recommendations.**
- **Time Trends and New Releases: Analysis of publication years suggests a bias toward newer books, but classic literature still holds strong**

appeal. A balanced inventory strategy that includes trending new releases and timeless classics can optimize long-term sales.

Business Recommendations:

1. **Regional Inventory Optimization: Align stock with regional demand, ensuring the most popular books are available in high-demand areas.**
2. **Targeted Marketing Campaigns: Promote hidden gems through personalized recommendations and digital campaigns.**
3. **Age-Based Book Recommendations: Implement recommendation engines in stores and online platforms based on customer demographics.**
4. **Data-Driven Pricing and Promotions: Leverage sales and rating trends to design discounts and loyalty programs around high-rated books.**
5. **Strategic Publisher Partnerships: Strengthen relationships with publishers of highly rated books to secure exclusive deals and early releases.**

By integrating these insights into business operations, bookstores can enhance customer satisfaction, increase sales, and maintain a competitive edge in the global market.