

ADVANCED DATA VISUALISATIONS TECHNIQUES USING AMAZON BOOKS DATASET



SECTION 1



Introduction: This comprehensive report provides insights into book genres, pricing trends, and author impacts by utilizing advanced analytics and visualization, we find out the contents of the dataset, offering strategic perspectives for readers.



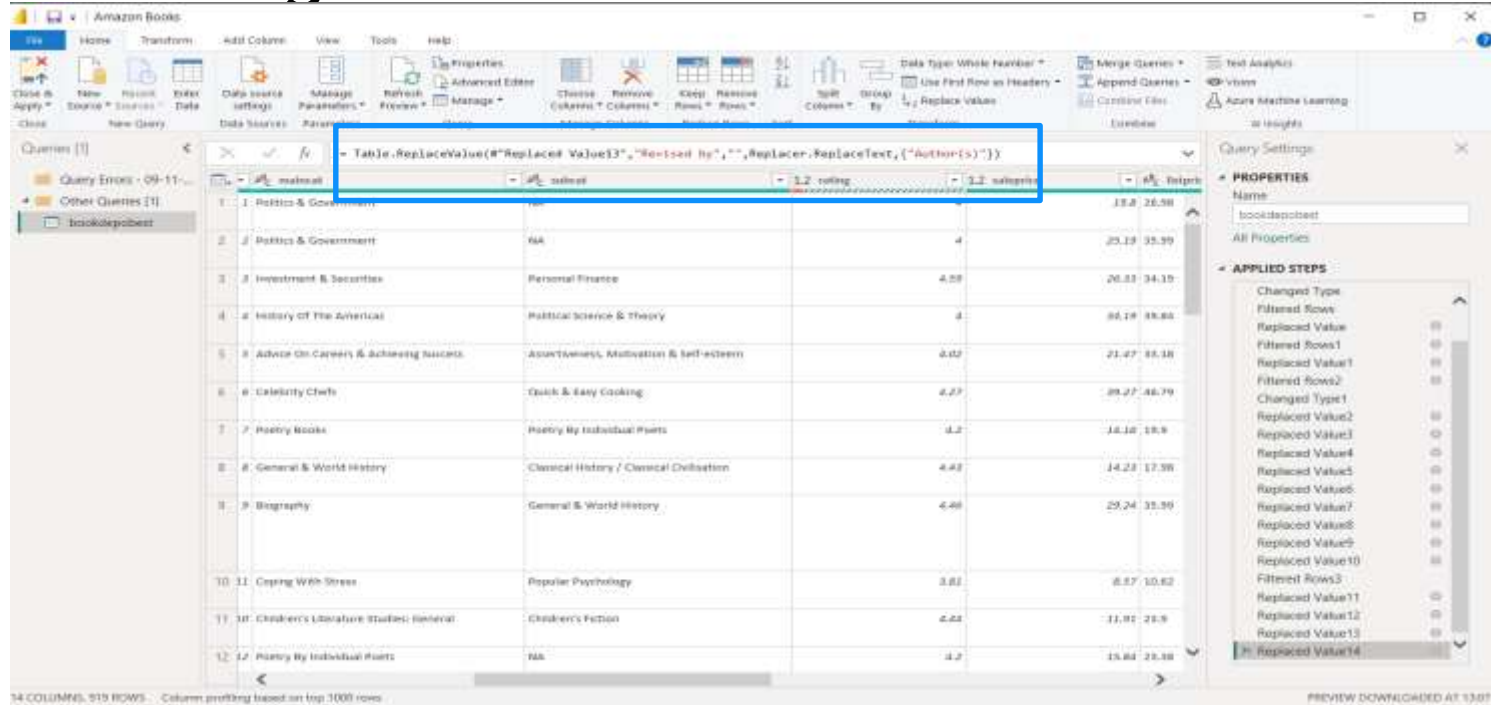
Link to Data: [Amazon Books Dataset](#)



Rationale for Choice: The dataset about Amazon products seems interesting as it contains diverse information such as product details, reviews, pricing, and categories. Analyzing this data could offer insights into customer preferences, and popular products, and potentially contribute to marketing strategies.



Handling Data Issues and Transformations:



Amazon Books

Table.ReplaceValue(#'Replaced Value13', 'Revised by', '', Replacer.ReplaceText, {'Author(s)'})

Rank	maincat	authors	goodreadsratecount	list price	rating
1	1. Politics & Government	NA	4	13.99	35.39
2	2. Politics & Government	NA	4	23.19	35.39
3	3. Investment & Securities	Personal Finance	4.59	26.53	34.10
4	4. History Of The Americas	Political Science & Theory	4	36.19	35.86
5	5. Advice On Careers & Achieving Success	Adults/Young, Motivation & Self-esteem	4.02	21.47	33.18
6	6. Celebrity Chefs	Quick & Easy Cooking	4.23	39.27	46.79
7	7. Poetry Books	Poetry By Individual Poets	4.2	14.16	19.9
8	8. General & World History	Classical History / Classical Civilisation	4.43	14.23	17.36
9	9. Biography	General & World History	4.46	29.24	35.99
10	10. Coping With Stress	Popular Psychology	3.82	6.57	10.62
11	11. Children's Literature Studies: General	Children's Fiction	4.64	11.91	25.9
12	12. Poetry By Individual Poets	NA	4.2	15.84	23.58

14 COLUMNS, 315 ROWS - Column profiling based on top 1000 rows

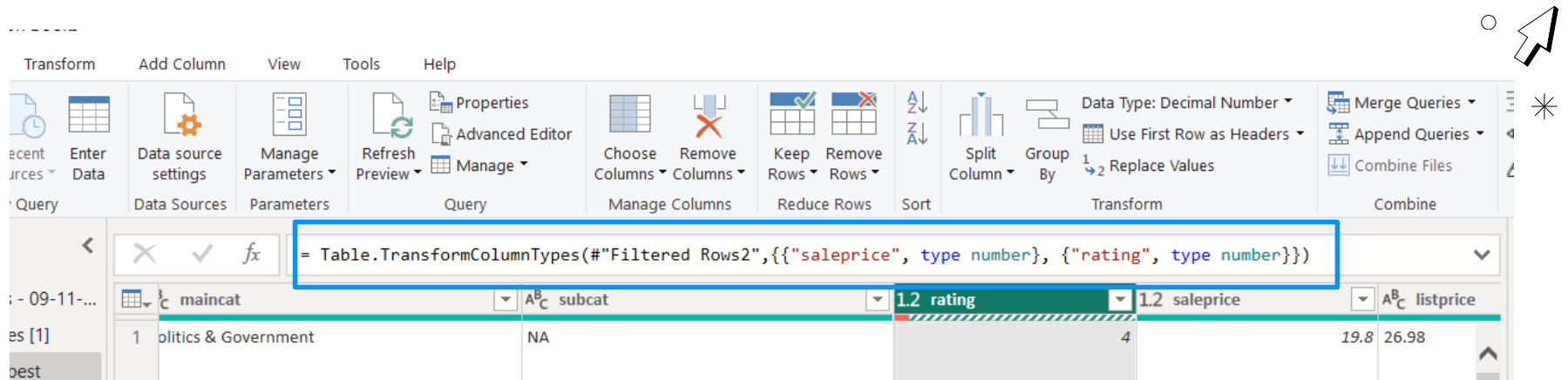
PREVIEW DOWNLOADED AT 13:07

Data Cleaning:

Missing Data: Identified missing data using queries in crucial columns such as 'saleprice', 'maincat', 'authors', 'goodreadsratecount', 'list price', and 'rating'.

Handling Null Values:

Replaced and removed missing values cautiously to ensure the dataset's integrity.



Transform Add Column View Tools Help

Recent sources Enter Data Data source settings Manage Parameters Refresh Preview Properties Advanced Editor Manage Query Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Decimal Number Use First Row as Headers Replace Values Merge Queries Append Queries Combine Files Combine

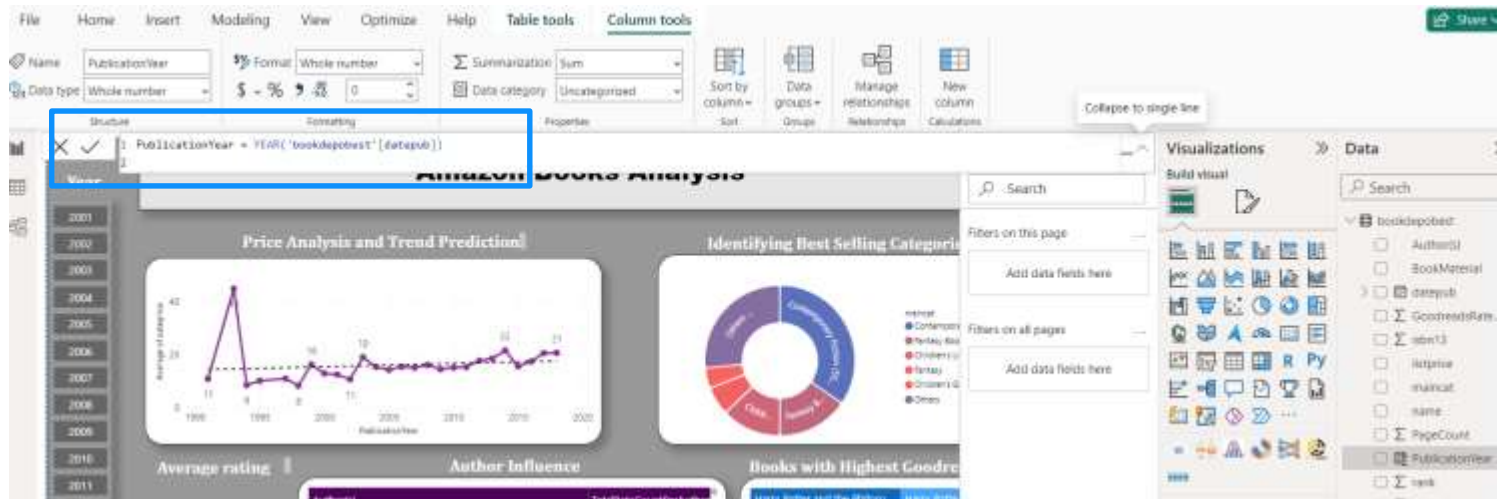
Formula Bar: `= Table.TransformColumnTypes("#Filtered Rows2",{{"saleprice", type number}, {"rating", type number}})`

	maincat	subcat	rating	saleprice	listprice
1	Politics & Government	NA	4	19.8	26.98

Data Transformation:

Convert Data Types: Converted string values to numerical formats for columns like 'sale price', 'list price', and 'rank', and converted string values into decimal numbers for the column 'rating'. Ensured that data types are appropriate for analysis.

Also, Converted string-based columns ('datepub') to the date format.



Feature Engineering:

Extracted useful features from existing data, such as the year from the publication date. By creating new column using DAX query: “`PublicationYear = YEAR('bookdepobest'[datepub])`”

Normalization:

Normalized numerical data ('saleprice', 'listprice') for better analysis using DAX query: “`ScaledListPrice = (bookdepobest[listprice] - MIN(bookdepobest[listprice])) / (MAX(bookdepobest[listprice]) - MIN(bookdepobest[listprice]))`”.



Selected Advanced Data Analytics Technique/Visual

Investigative Questions:

Identify Bestselling Categories:

- Analysis Method: Use Power Query and DAX expressions for aggregating total reviews or Goodreads rate count by the main category.

Price Analysis and Trend Prediction:

- Analysis Method: Trendlines to analyze price trends over publication years.

Author Influence:

- Analysis Method: Clustering techniques or Decision Tree in Power BI to categorize or segment authors based on ratings, rate counts, or sales.



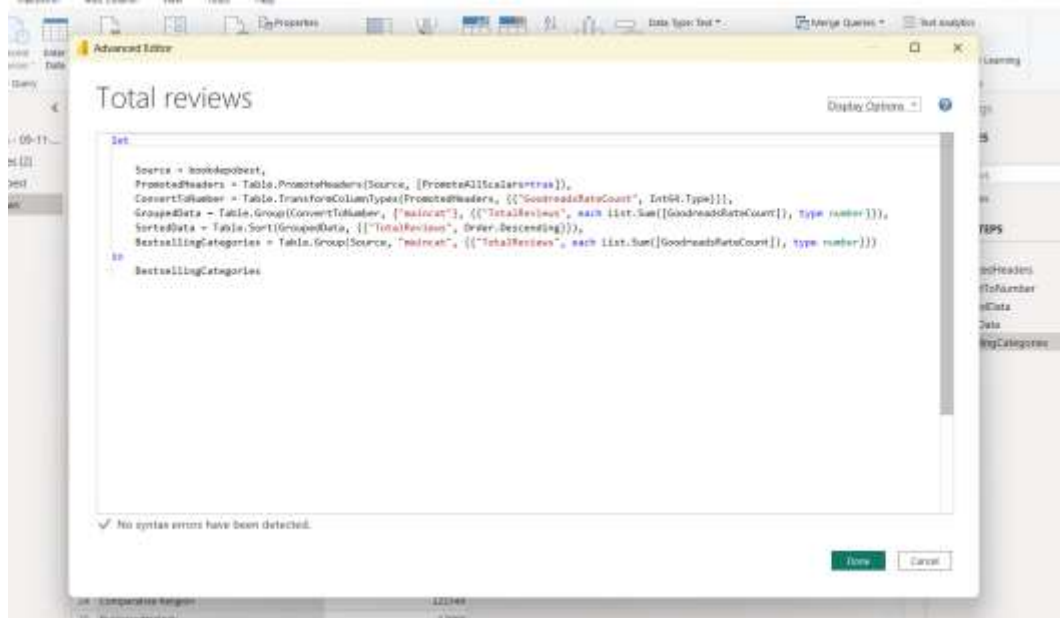
○

Identify Bestselling Categories:

✱

- Analysis Method: Use Power Query and DAX expressions for aggregating Goodreads rate count by the main category.

Used this query to group the dataset by the main category, calculate the total reviews for each category, sort the data in descending order based on total reviews, and then select the top category as the bestselling category.

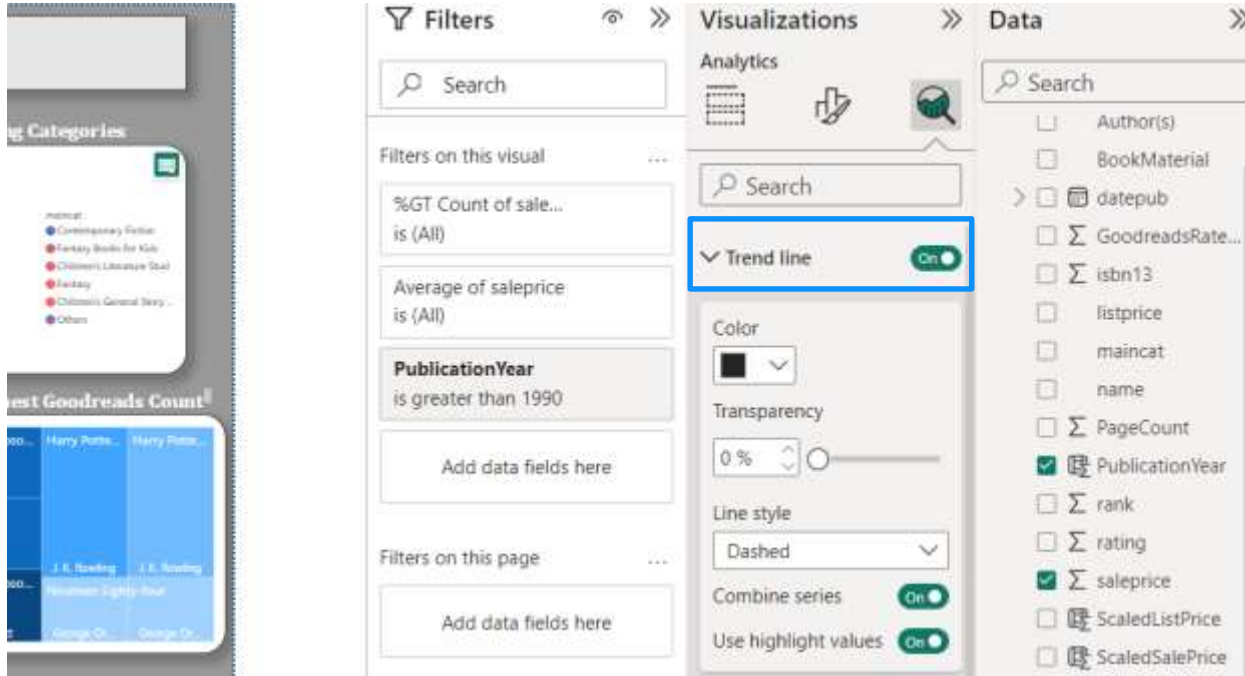


✱

○ Price Analysis and Trend Prediction:

- * ■ Analysis Method: Trendlines to analyze price trends over publication years.

Used the 'Analytics' pane to add a 'Trend Line' to the chart that display a trendline



○

Author Influence:

- * ■ Analysis Method: Clustering techniques to categorize authors based on ratings, rate counts, or sales.

Used the following DAX Expression for Clustering Authors into new a table:

The screenshot shows the DAX editor for a measure named 'TotalRateCountPerAuthor'. The formula bar contains the following DAX expression:

```
1 TotalRateCountPerAuthor =
2 SUMX(
3     VALUES('bookdepobest'[Author(s)]),
4     CALCULATE(SUM('bookdepobest'[GoodreadsRateCount]))
5 )
```

The interface includes tabs for 'Structure', 'Formatting', and 'Properties'. The 'Formatting' tab is active, showing the 'Format' dropdown set to 'Whole number' and the 'Data category' dropdown set to 'Uncategorized'.

Total rate count per Author

The screenshot shows the DAX editor for a measure named 'TotalSalesPerAuthor'. The formula bar contains the following DAX expression:

```
1 TotalSalesPerAuthor =
2 SUMX(
3     VALUES('bookdepobest'[Author(s)]),
4     CALCULATE(SUM('bookdepobest'[saleprice]))
5 )
```

The interface includes tabs for 'Structure', 'Formatting', and 'Properties'. The 'Formatting' tab is active, showing the 'Format' dropdown set to 'Whole number' and the 'Data category' dropdown set to 'Uncategorized'.

Total sales per Author

The screenshot shows the DAX editor for a measure named 'AvgRatingPerAuthor'. The formula bar contains the following DAX expression:

```
1 AvgRatingPerAuthor =
2 AVERAGEX(
3     VALUES('bookdepobest'[Author(s)]),
4     CALCULATE(AVERAGE('bookdepobest'[rating]))
5 )
```

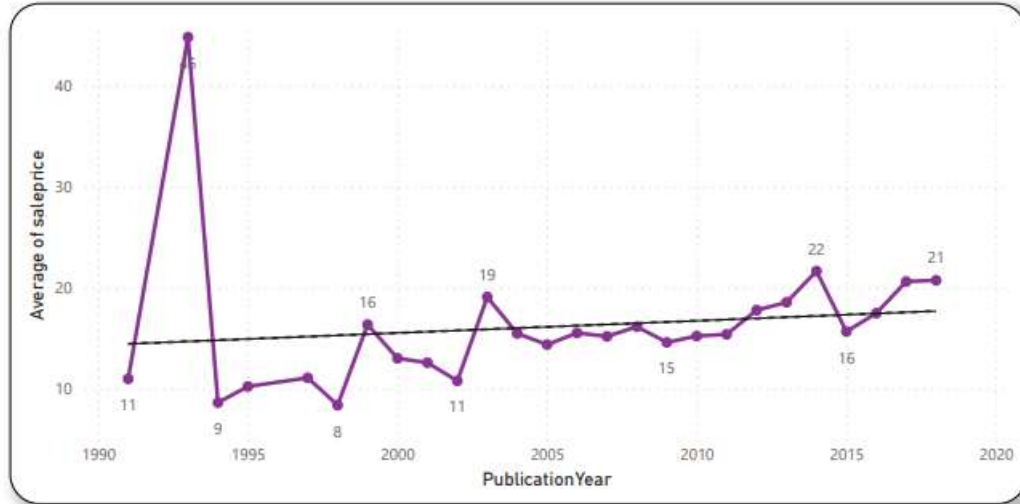
The interface includes tabs for 'Structure', 'Formatting', and 'Properties'. The 'Formatting' tab is active, showing the 'Format' dropdown set to 'Whole number' and the 'Data category' dropdown set to 'Uncategorized'.

Average rating per Author

✱



Section 3: Visualizations (Diagrams)



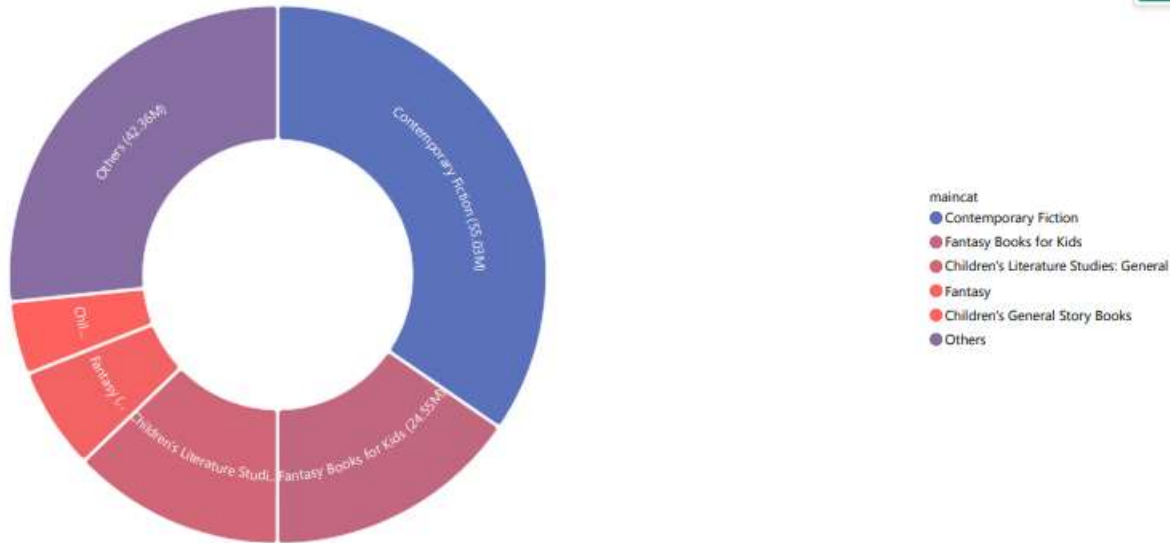
Price Analysis and Trend Prediction:

Objective: Analyze pricing trends over the years using the trend line method

Visualization: Line Chart with Trendline

At 44.81, 1993 had the highest Average sale price and was 433.77% higher than 1998, which had the lowest Average sale price at 8.40. Across all 26 Publication Years, the Average sale price ranged from 8.40 to 44.81.

Bestselling Categories



Objective: Identify Bestselling Categories

Method: Power Query and DAX Expressions

Visualization: Sunburst Chart

Author Influence

Author(s)	TotalRateCountPerAuthor	TotalSalesPerAuthor	AvgRatingPerAuthor
J. K. Rowling	44946323	559	4
J. K. Rowling , Jim Kay	8894756	143	4
George Orwell	6260293	34	4
J. K. Rowling , Matthew Fitt	4991236	11	4
Harper Lee	3376674	11	4
John Steinbeck , Susan Shillinglaw	3070532	20	4
Paulo Coelho	2864335	43	4
F. Scott Fitzgerald , Guy Reynolds , Dr. Keith Carabine	2862324	5	4
Lois Lowry	2722272	21	4
William Shakespeare , Roma Gill	2217096	22	4
J. D. Salinger	2146445	11	4
George Orwell , Thomas Pynchon	2095625	14	4
George Orwell , Malcolm Bradbury	2069043	11	4
Total	159115238	16839	4

Objective: Identify the influence of authors

Method: Clustering Techniques

Visualization: Table



Visualizations Overview

Line chart

Highest Average Sale Price:

- Year 1993: 44.81

Lowest Average Sale Price:

- Year 1998: 8.40

Table chart

Top Authors by Total Rate
Count:

- J. K. Rowling: 44,946,323
- J. K. Rowling Jim Kay: 8,894,756
- George Orwell: 6,260,293
- J. K. Rowling, Matthew Fitt: 4,991,236
- Harper Lee: 3,376,674

Top Authors by Total Sale
value:

- J. K. Rowling: 559
- J. R. R. Tolkien: 216
- Brandon Sanderson: 208
- Moleskine: 180
- Lonely Planet: 179



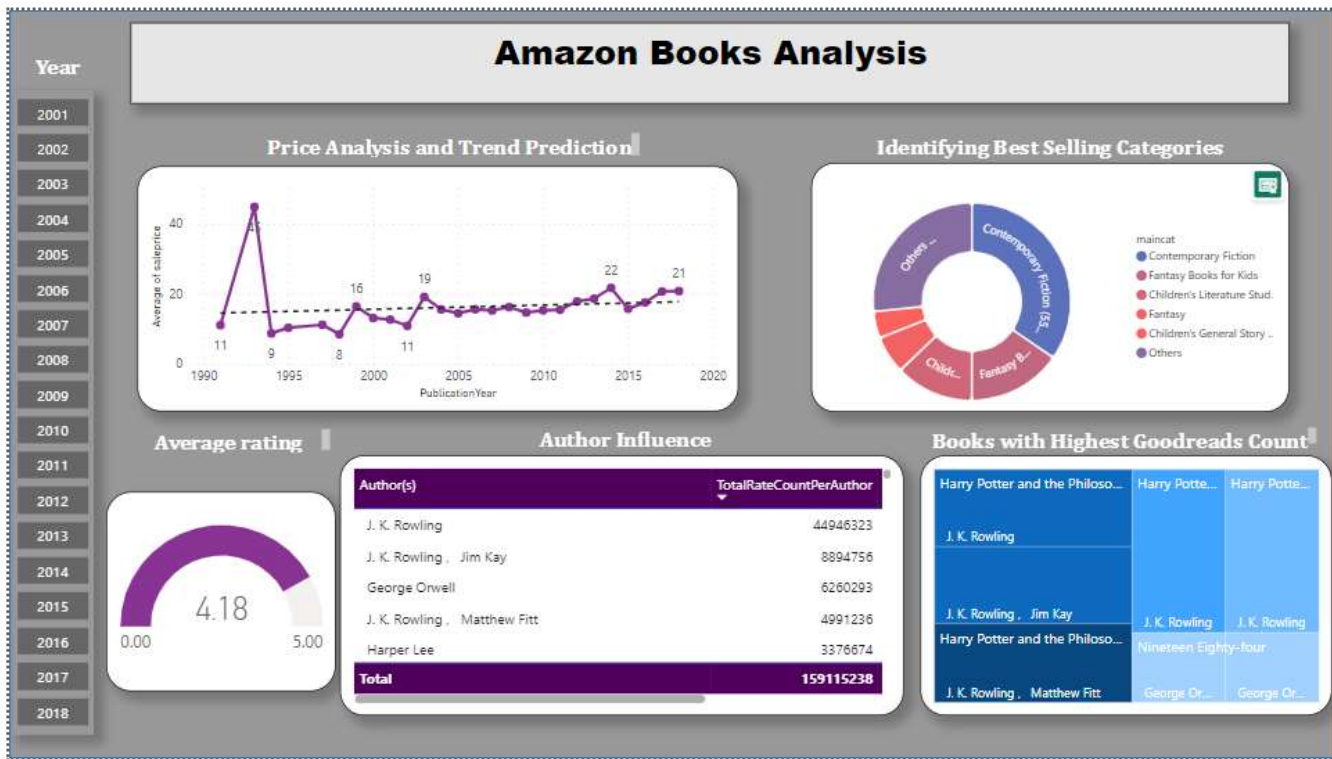
Visualizations Overview

Sun burst chart

- Contemporary Fiction: 550.32M (Highest) - Strong engagement in contemporary fiction with over half a billion good reviews.
- Fantasy Books for Kids: 245.54M - Robust reviews indicate high popularity in the fantasy genre for children.
- Children's Literature Studies: General: 203.80M - Strong academic interest in children's literature studies reflected by millions of good reviews.
- Fantasy: 98.14M - The fantasy genre maintains a substantial following with nearly 100 million good reviews.

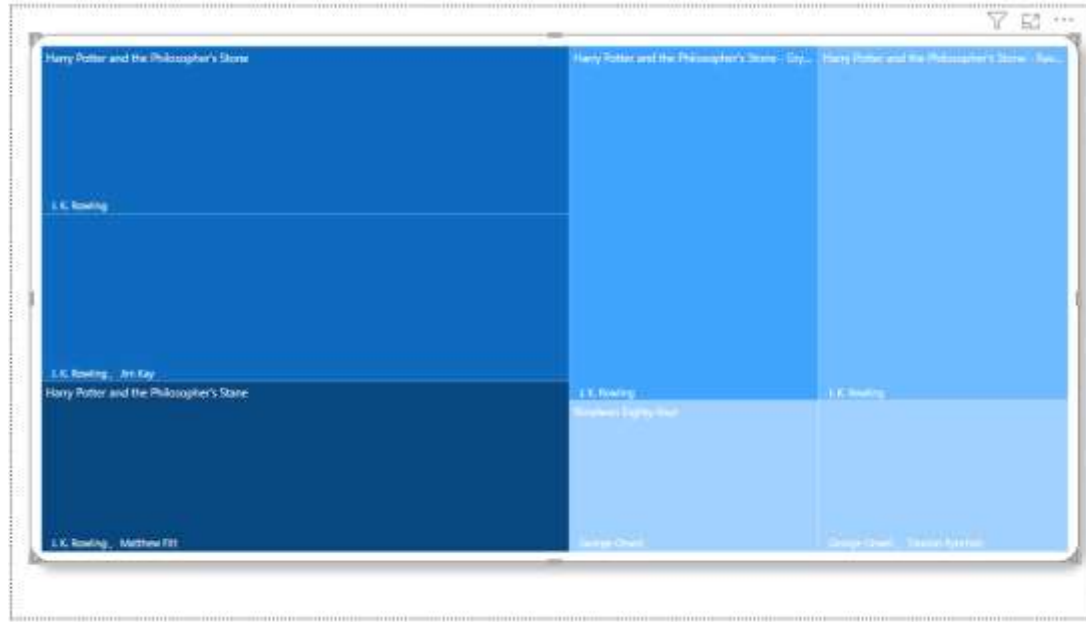


Dashboard



The dashboard provides valuable insights into pricing trends, popular book categories, customer satisfaction levels, and influential authors, aiding in strategic decision-making for the Amazon book market.

Tree map



The treemap visualization indicates that "Harry Potter and the Philosopher's Stone" by J.K Rowling has the highest Goodreads RateCount with an impressive count of 4,991,236.

Section4: Findings and Limitations

Findings

- Genre Popularity
- Individual Book Impact
- Pricing Trends
- Author Influence

Limitations

- Data completeness
- Limited of external factors
- Limited Book Details
- Genre Specifies



Conclusion



- The results of our analysis illustrate the continuing popularity of children's literature studies, fantasy, and contemporary fiction. With an amazing Goodreads RateCount, "Harry Potter and the Philosopher's Stone" becomes an incredible literary masterpiece. Forecasted pricing trends offer useful information for making strategic pricing decisions.
- The results improve our knowledge of author influences and genre dynamics despite limitations in data completeness and external factors

References

- Chola', A. (2023). DAX Functions in Power BI: 8 Popular Function Types. Retrieved from hevota.com.
- Microsoft. (2023). Advanced Analytics . Retrieved from arbellatech.com.
- Microsoft. (2023). Create custom data visualisations for Power BI. Retrieved from powerbi.microsoft.com.
- pbivizedit. (2023). Advanced Line Chart. Retrieved from [pbivizedit](https://pbivizedit.com).