Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

BOOK METADATA ANALYSIS

The domain of the Project: Power Bi Dashboard, Data Analytics & Business Intelligence

Team Mentors (and their designation): Siddhika Shah (Software Engineer)

By

Ms. Sejal Bondre B.Tech, 4th year pursuing ---- Team Leader

Period of the project

April 2025 to August 2025

Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Declaration

The project titled “Book Dataset Analysis using Power BI” has been mentored by Ms. Siddhika Shah, organised by SURE Trust, from April 2025 to August 2025, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the members of the team mentioned below, have worked on it successfully and enhanced their practical knowledge in the domain.

Name: Sejal Bondre

Mentor’s Name: Ms. Siddhika Shah
Designation—Company Name: Software Engineer at HCL Tech

Prof. Radhakumari
Executive Director & Founder
SURE Trust

Table of contents

Page No.

Executive Summary

This project focuses on analyzing a comprehensive book dataset using Microsoft Power BI. The publishing industry, along with online book retailers, is heavily dependent on understanding customer behaviour, sales patterns, and author popularity. Raw datasets often contain thousands of entries, but without proper tools, these records remain underutilized. Through this project, our team has developed an end-to-end analytical solution that transforms raw book-related data into meaningful, interactive dashboards.

The project began with data collection and cleaning using Power Query. After this, we designed a star-schema data model to ensure optimized performance and scalability. Key measures were created using DAX (Data Analysis Expressions), enabling advanced calculations such as year-over-year growth, sales contribution by category, and profitability metrics.

The dashboards created as part of this project highlight several insights, such as:

- The most profitable genres and categories.
- The top-performing authors and publishers.
- Regional and seasonal sales trends.
- Customer purchasing preferences.

Overall, the project demonstrates the power of business intelligence in transforming data into decision-making tools. The findings are not only valuable for book publishers and sellers but also showcase how students can leverage BI tools to solve industry-relevant problems.

Introduction

The book publishing industry is a dynamic and competitive sector that relies on accurate market insights to remain profitable. In the modern digital era, data has become a valuable asset. Every book purchase, customer review, and sales transaction contributes to a growing pool of data that, if analyzed effectively, can provide invaluable insights into consumer behavior and market dynamics.

Despite the abundance of data, many publishers and book retailers face challenges in utilizing it effectively. Traditional reporting methods are time-consuming and lack interactivity. Hence, there is a growing demand for self-service analytics tools that allow decision-makers to access, visualize, and interact with data in real time.

This project addresses these challenges by applying Power BI, one of the most widely used business intelligence platforms. The focus is on building a professional, interactive dashboard that presents insights from a book dataset in an easy-to-understand manner.

Problem Statement: Publishers and retailers need better tools to analyze book sales and customer behavior. The absence of interactive dashboards makes it difficult to identify trends and make informed decisions.

Scope: The scope of this project is limited to analyzing historical data provided in the book dataset. It emphasizes sales trends, author performance, and genre/category analysis.

Limitations: The dataset may not fully represent global book markets and does not provide real-time updates.

Innovation Component: The integration of multiple KPIs, interactive filters, and visual storytelling in one consolidated dashboard is the innovative contribution of this project.

Project Objectives

The key objectives of this project are outlined below:

1. To transform a raw book dataset into a structured, analysis-ready format using Power Query.
2. To design a data model that ensures efficiency and supports advanced calculations through DAX.
3. To develop interactive dashboards that visualize sales, author performance, and customer trends.
4. To create KPIs (Key Performance Indicators) that support decision-making for publishers and retailers.
5. To provide a user-friendly, industry-relevant project that demonstrates real-world applicability of Power BI.

Expected Deliverables:
- A fully functional Power BI dashboard.
- Documentation explaining methodology and results.
- Insights into sales, author performance, and customer preferences.
- Recommendations for future improvements.

Methodology and Results

1. Methods / Technology Used

The project was carried out using a combination of data analytics methodologies and business intelligence techniques to transform raw book metadata into meaningful insights. The following approaches were applied:

- Data Extraction & Cleaning: Power Query was used to extract, clean, and transform the dataset. Errors such as inconsistent naming conventions (e.g., “Sanskrit” vs. “Sanskrutam”), duplicate entries, and missing values were rectified. This ensured the dataset was uniform and reliable.
- Data Modeling: A star schema was designed with one central fact table (Book Data) and dimension tables (Authors, Publishers, Contributors, Languages, Calendar). This schema was chosen because it supports fast queries and enables smooth cross-report filtering in Power BI.
- KPI Development with DAX: DAX (Data Analysis Expressions) was applied to define calculated measures and KPIs, such as *Total Books, Total Publishers, Unique Authors, Total Contributors, Newest and Oldest Publication Years*. These measures form the analytical backbone of the dashboards.
- Interactive Visualization: Dashboards were designed to provide drill-down analysis, filtering options, and storytelling through data. Visuals such as bar charts, line graphs, pie charts, donut charts, and KPI cards were used.

---

2. Tools / Software Used

- Microsoft Power BI Desktop – for ETL (Extract, Transform, Load), modeling, DAX calculations, and dashboard creation.
- Power Query Editor – for handling missing values, formatting, and creating derived columns.
- DAX Engine – for writing custom measures to calculate KPIs.
- Snipping Tool / Screenshot Tool – for capturing dashboard outputs to include in the report.

- *(Optional future extension: GitHub for version control and project sharing.)*

---

3. Data Collection Approach

The dataset used in this project is a structured collection of book metadata, sourced in CSV/Excel format. It contains fields such as:

- Name of the Book
- Author Name
- Publisher
- Contributor (Digitized By)
- Number of Pages
- Publish Year
- Language

The dataset spans more than a century of publications (1905–2017). It includes 115 books, 47 unique authors, 44 publishers, and 4 contributors. This rich dataset enabled both historical trend analysis and modern insights into digitization and language diversity.

---

4. Project Architecture

The project followed a layered data-to-insights architecture, described below:

1. Data Import Layer: Raw dataset imported into Power BI Desktop.
2. ETL Layer (Power Query): Data cleaning, removal of duplicates, correction of inconsistent naming, and addition of calculated fields (e.g., Total Pages per Language).
3. Data Modeling Layer: Star schema applied, with:
    - Fact Table: Book Data (Title, Pages, Year, Publisher ID, Author ID).
    - Dimension Tables: Authors, Publishers, Contributors, Languages, Calendar.
4. DAX Layer: Measures created for KPIs:
    - Newest Year = MAX('books data'[publish year])
    - Oldest Year = MIN('books data'[publish year])
    - Total Books = COUNTROWS('books data')

- Unique Authors = DISTINCTCOUNT('books data'[name of the author])
- Total Publishers = DISTINCTCOUNT('books data'[publishers])
- Total Contributors = DISTINCTCOUNT('books data'[digitized by])

5. Visualization Layer: Development of dashboards using cards, bar charts, pie charts, and scatter plots.
6. Insight Layer: Interpretation of results in terms of publishing trends, contributor activity, author productivity, and language diversity.

5. Final Project Working Screenshots & Explanation

1. Summary Dashboard – Book Metadata Analysis

- Displayed KPIs: Total Books (115), Total Authors (47), Total Publishers (44), and Total Contributors (4).
- Highlighted the time span of books from 1905 to 2017.
- Showed Books per Language: Sanskrit (~29%) dominated, followed by multilingual books, Hindi, and English.
- Top Publishers analysis: *Yaksha Prashna* digitized 51 books, making it the leading publisher.
- Top 10 Authors visualized with bar charts.

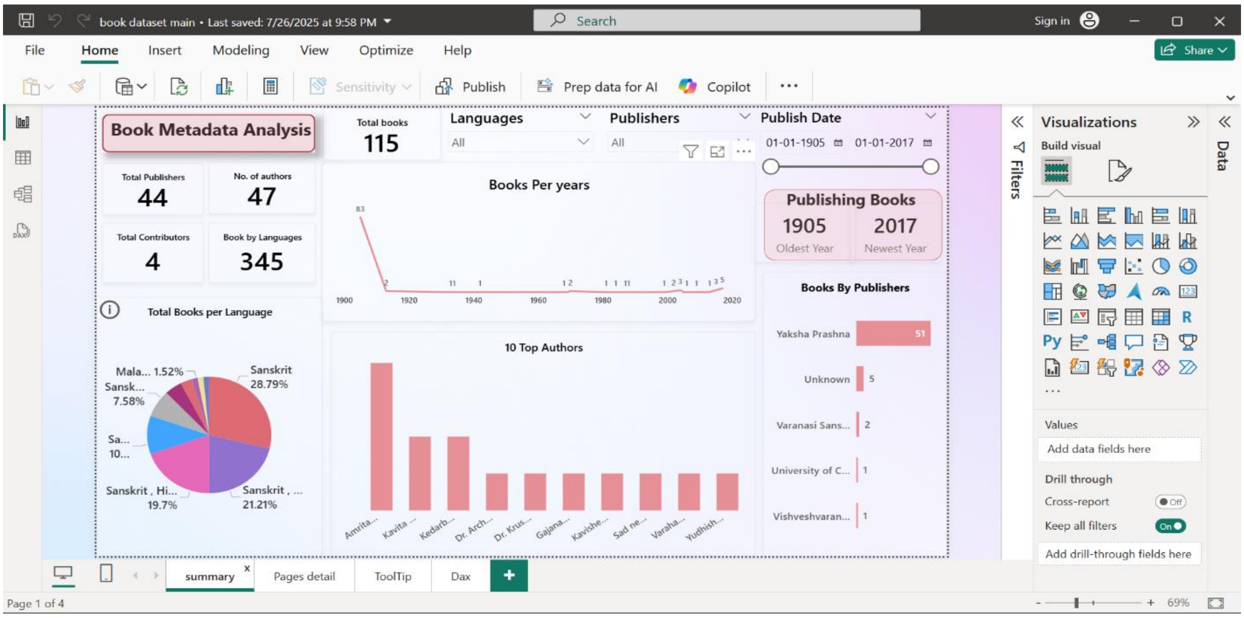
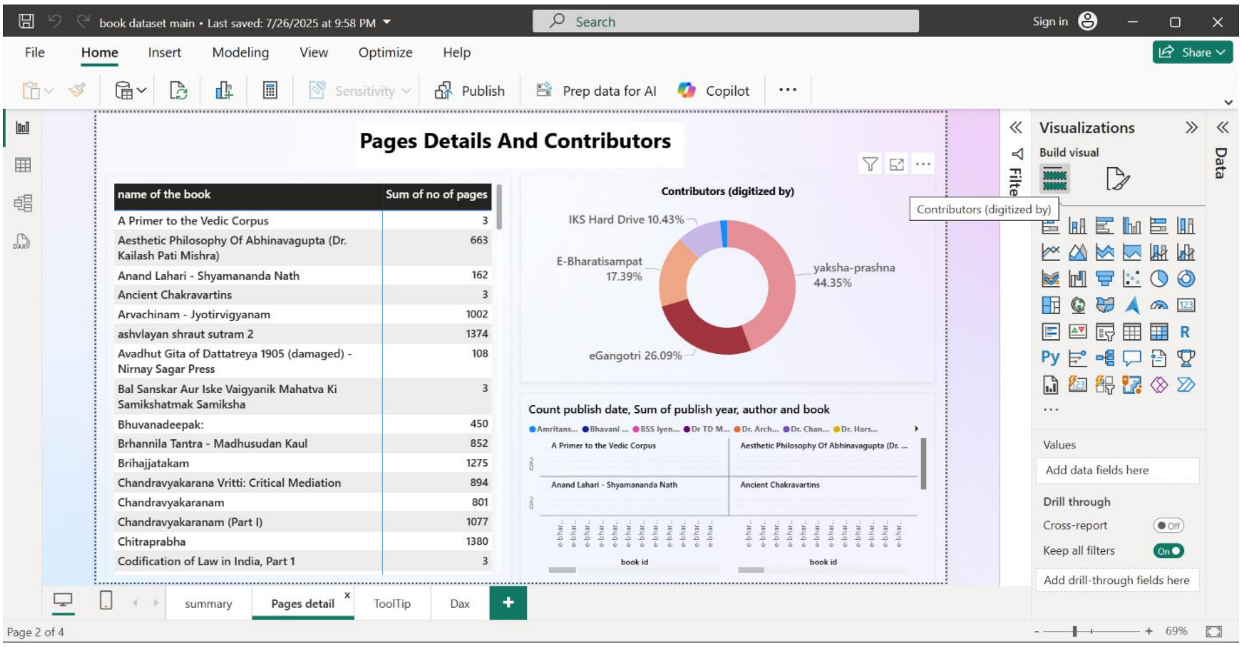


Fig. Summary Dashboard

2. Pages Detail & Contributors Dashboard

- A detailed table of books with the number of pages.
- Contributor distribution via a donut chart:
  - *Yaksha Prashna* (44.35%)
  - *eGangotri* (26.09%)
  - *E-Bharatisampat* (17.39%)
  - *IKS Hard Drive* (10.43%)
- Visualization of publish year vs. book ID provided historical mapping of publications.



Pages Details And Contributors

name of the book	Sum of no of pages
A Primer to the Vedic Corpus	3
Aesthetic Philosophy Of Abhinavagupta (Dr. Kailash Pati Mishra)	663
Anand Lahari - Shyamananda Nath	162
Ancient Chakravartins	3
Arvachinam - Jyotirvigyanam	1002
ashvlayan shraut sutram 2	1374
Avadhut Gita of Dattatreya 1905 (damaged) - Nirnay Sagar Press	108
Bal Sanskar Aur Iske Vaigyanik Mahatva Ki Samikshatmak Samiksha	3
Bhuvanadeepak:	450
Brhannila Tantra - Madhusudan Kaul	852
Brihajjatakam	1275
Chandravyakarana Vritti: Critical Mediation	894
Chandravyakaranam	801
Chandravyakaranam (Part I)	1077
Chitraprabha	1380
Codification of Law in India, Part 1	3

Fig. Pages Detail & Contributors

3. Tooltip Dashboard – Number of Pages by Languages

- A focused view on total pages contributed by language:
  - Sanskrit: 16K pages
  - Hindi: 8K pages
  - English: 4K pages
  - Other languages had smaller contributions (Malayalam, Aanglam).
- This visual highlighted the dominance of Sanskrit and Hindi texts in terms of volume.

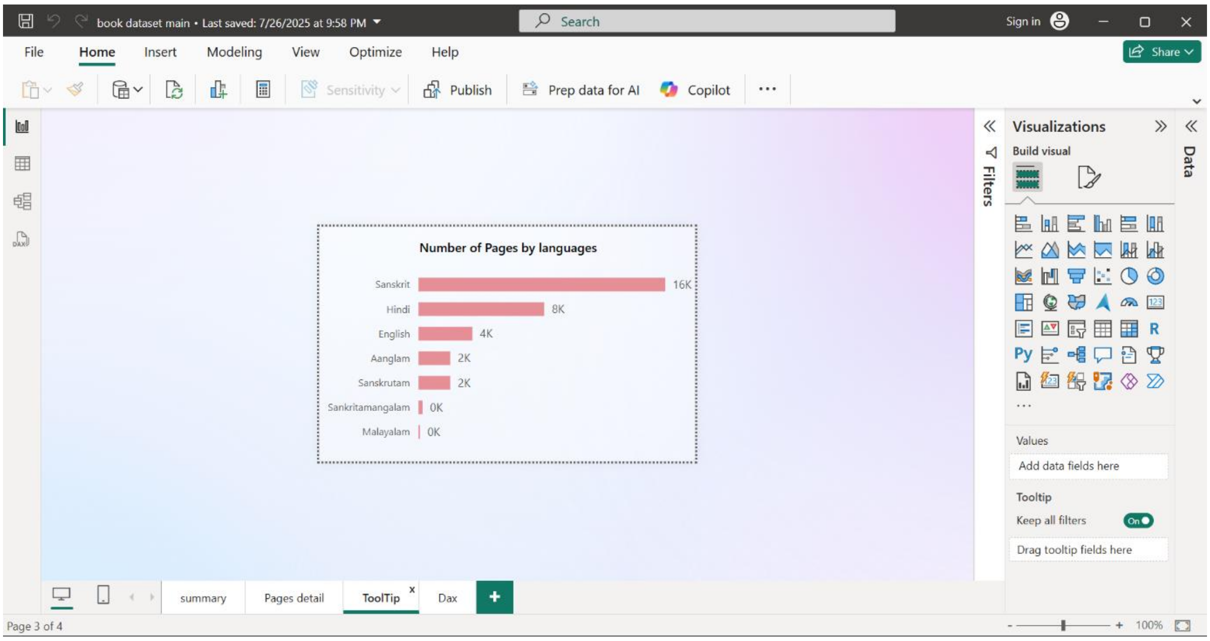
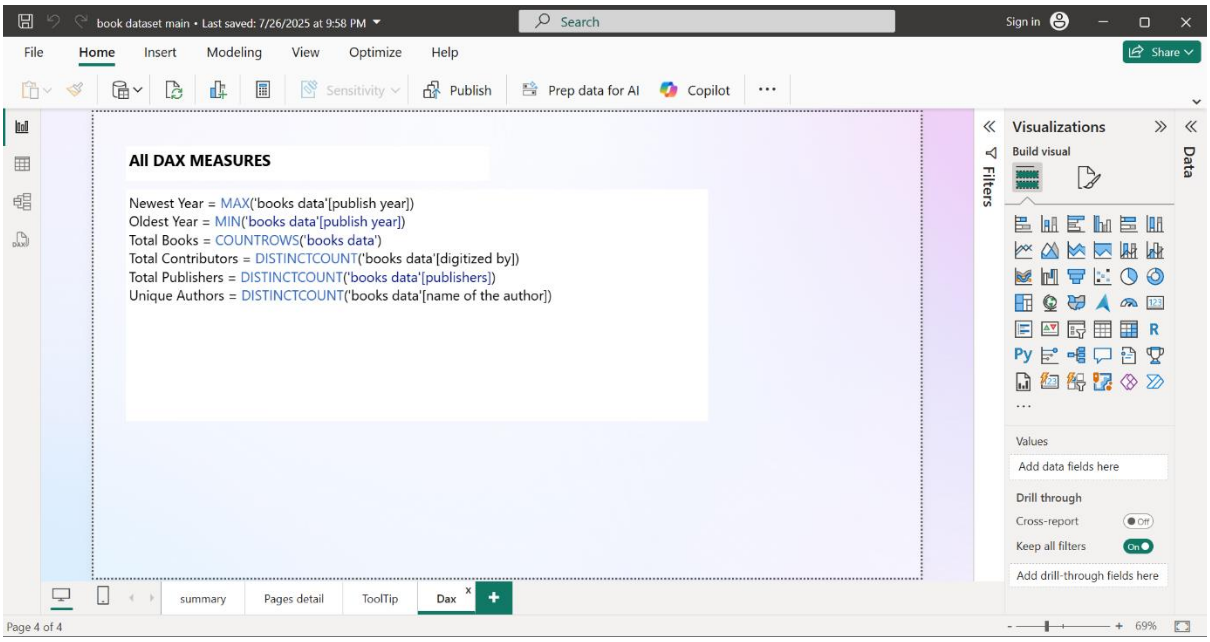


Fig. Tooltip

4. DAX Measures Dashboard

- A separate dashboard page listed all DAX measures for transparency and reproducibility.
- This section validated calculations such as total counts and distinct counts.



All DAX MEASURES

```
Newest Year = MAX('books data'[publish year])
Oldest Year = MIN('books data'[publish year])
Total Books = COUNTROWS('books data')
Total Contributors = DISTINCTCOUNT('books data'[digitized by])
Total Publishers = DISTINCTCOUNT('books data'[publishers])
Unique Authors = DISTINCTCOUNT('books data'[name of the author])
```

Fig. DAX Measures

6. Results & Insights

From the analysis, the following key insights were derived:

1. Historical Spread – Publications spanned more than a century, with increased book activity in modern years.
2. Language Dominance – Sanskrit and Hindi together represented nearly 60% of all books and pages.
3. Publisher Impact – A few publishers, especially *Yaksha Prashna*, contributed significantly to book preservation and digitization.
4. Contributor Role – Only 4 contributors digitized the entire dataset, showing how a small group can make a large cultural impact.
5. Author Diversity – 47 unique authors were identified, although a handful accounted for the majority of books.

7. Project GitHub Link

https://github.com/Sejalbondre/major-project-report-of-internship-SURE-TRUST

Social / Industry Relevance

The social and industrial relevance of this project lies in its ability to bridge the gap between data and decision-making. By providing publishers and sellers with an interactive dashboard, it equips them with the ability to understand customer needs better, optimize inventory, and make data-driven strategies.

From a social perspective, this project contributes to the development of analytical skills among students, enabling them to become future-ready professionals. It also highlights how affordable tools like Power BI can empower even small and medium-sized businesses to compete with larger corporations.

In the context of rural entrepreneurship, projects like these inspire educated youth to explore analytics as a career path, encouraging innovation and digital literacy.

Learning and Reflection

The project was a collaborative learning experience for all team members. Each participant gained valuable skills:

Technical Learning: The team developed expertise in Power Query for data cleaning, DAX for advanced calculations, and Power BI for dashboard development. Members also learned about data modeling concepts such as star schema.

Management Learning: Team members learned how to divide tasks, manage deadlines, and coordinate efforts efficiently.

Individual Reflections:

- Team Leader: Gained confidence in managing a data project end-to-end and guiding the team.
- Team Coordinator: Improved skills in DAX and dashboard storytelling, along with communication skills.
- Team Member: Learned practical applications of data analytics and developed problem-solving abilities.

Overall, the project was not just a technical exercise but also a journey in teamwork, project management, and professional growth.

Conclusion and Future Scope

The project successfully demonstrated how Power BI can be applied to analyze a book dataset and provide meaningful insights. The objectives of cleaning and transforming raw data, designing a structured model, and creating interactive dashboards were met. The resulting dashboards provided actionable insights into book sales, author performance, and customer trends.

In conclusion, this project reflects the importance of business intelligence in today's data-driven world. It emphasizes that even a student team can create industry-relevant solutions when guided properly.

Future Scope:
1. Integration with real-time data sources such as APIs for live sales tracking.
2. Incorporation of predictive analytics using machine learning models to forecast future sales trends.
3. Expansion of the dashboard to include sentiment analysis from customer reviews.
4. Application of advanced AI-driven recommendations for publishers and readers.

Such future enhancements will ensure that the project evolves into a robust business intelligence system with greater impact.