INTRODUCTION TO DATA MANAGEMENT

PROJECT REPORT

(Project Semester January-April 2025)

LIBRARY BOOKS VISUALIZATION

Submitted by

Atul Kumar

Registration no:

12313543

Section: K23SK

Course Code: INT217

Under the Guidance of

Aashima

Assistant Professor

(28968)

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

<u>CERTIFICATE</u>

This is to certify that **Atul Kumar** bearing Registration no. **12313543** has completed   **INT217** project titled,

"Library Books Visualization" under my guidance and supervision.

 To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Aashima Assistant Professor School of Computer Science Engineering
Lovely Professional University Phagwara,

Punjab.

Date: 9th April, 2025

DECLARATION

I, <u>Atul Kumar</u>, student of B.tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 9th April, 2025   Signature

Registration No.12313543   Atul Kumar

Acknowledgment

I would like to express my deepest gratitude to Prof. Aashima for his exceptional mentorship and unwavering support throughout the duration of this project. His vast knowledge in the fields of data science and machine learning, combined with his patient and thoughtful guidance, played a pivotal role in the successful completion of this work. His insightful suggestions and feedback consistently challenged me to think critically and improve the quality of my research. I am also grateful for the learning environment he fostered, which encouraged exploration and innovation.

In addition, I sincerely thank my peers and classmates for their helpful discussions, encouragement, and collaborative spirit during this project. Their input provided fresh perspectives that contributed meaningfully to the final outcome. I am also thankful to the open-source community for providing the tools, libraries, and resources that made the implementation of this project possible. Lastly, I acknowledge the dataset contributors for making this analysis feasible.

TABLE OF CONTENT:

1. Introduction

The management and analysis of library book datasets are absolutely essential for ensuring efficient tracking, resource allocation, and strategic decision-making in educational institutions, public libraries, and research centers. This report is dedicated to a thorough and meticulous analysis of a comprehensive dataset that contains detailed information about books, including their titles, authors, genres, publishers, availability status, publication years, editions, shelf locations, costs, categories, book conditions, number of reviews, and barcodes. The dataset serves as a rich and valuable source of information, providing deep insights into publishing trends, book usage patterns, and user preferences based on review counts and physical conditions.

The primary goal of this report is to delve deeply into the dataset to extract meaningful, actionable, and insightful information that can greatly assist librarians, educators, administrators, and other stakeholders in managing library resources more effectively and efficiently. We will address five specific and well-defined objectives in this analysis: (1) determining in a detailed manner the exact number of books published by each publisher across different genres to understand publishing distribution, (2) enabling a precise search for all comprehensive details of any specific book by using its unique Book ID, simulating a VLOOKUP-like functionality for quick access, (3) ranking all publishers based on the total number of reviews their books have received to assess their popularity and impact, (4) creating a detailed visualization of books organized by their shelf locations ranging from A1 to E4, utilizing interactive slicers to allow for dynamic filtering and exploration of the data, and (5) developing a fully functional system to generate formatted receipts for book transactions based on user and book details. These objectives are designed to provide a holistic understanding of the distribution, usage patterns, availability status, and overall popularity of books within the library system.

This analysis holds both academic significance and practical utility, as it can significantly contribute to optimizing library operations, enhancing user satisfaction through better resource availability, and informing future decisions regarding book acquisitions, shelf reorganization, and maintenance schedules. By carefully examining each objective, this report aims to deliver clear, actionable insights that can drive improvements in library management practices and user experience, ensuring that the library remains a valuable and efficient resource for all its patrons.

2. Source of Dataset

- **Real Time DataSets**

3. Dataset Preprocessing

Before embarking on the detailed analysis, the dataset underwent a rigorous preprocessing phase to ensure data quality, consistency, and reliability. The raw dataset initially comprised 7,405 rows and 17 columns, but it

presented several challenges, including missing values, inconsistent data formats, and potential errors or anomalies in certain fields such as ISBN and Barcode. For instance, some entries had missing values in fields like Issue Date and Return Date for books marked as "Available," while other fields exhibited inconsistencies, such as book conditions labeled variably as "NEW," "AVG," or "WORN" with occasional extra spaces or typos.

The preprocessing steps were carefully executed as follows: First, we conducted a thorough data cleaning process, which involved identifying and removing any duplicate Book IDs to avoid redundancy and checking for inconsistencies in the formats of ISBN and Barcode fields, such as standardizing the "*" prefixes and suffixes in Barcodes. Second, we addressed missing values by retaining blank Issue Date and Return Date fields for books that were marked as "Available," as this was logically consistent and did not require imputation. Third, we standardized naming conventions across the dataset, ensuring uniformity in publisher names (e.g., "NCERT," "Macmillan") and genre labels (e.g., "Maths," "Economics"), correcting any variations or errors like trailing spaces. Fourth, we converted data types where necessary, transforming numerical fields such as "Cost," "Pages," and "Number of Reviews" into numeric formats for accurate calculations, and standardizing date fields like Issue Date and Return Date into a consistent date format for easier analysis. Finally, we filtered the dataset to focus only on the subsets of data relevant to our specific objectives, such as genres, publishers, and shelf locations, ensuring that no irrelevant or extraneous data would skew our results.

After completing these preprocessing steps, the dataset was transformed into a clean, consistent, and reliable form, with 7,405 records ready for in-depth analysis, guaranteeing that the subsequent results would be both accurate and trustworthy.

4. Analysis on Dataset

Objective 1: Genre-Wise Publisher Counts

-        General Description: This objective focuses on a detailed examination of how many books each publisher has published for each specific genre, providing a clear picture of publishing distribution across different categories of books.
-        Specific Requirements: The task requires counting the exact number of books published by each publisher for every genre, ensuring a comprehensive breakdown that highlights which publishers dominate in which areas. - Analysis Results: The analysis revealed that NCERT published the highest number of books overall, totaling 1,258, closely followed by Oxford with 1,268 books. When broken down by genre, NCERT emerged as the leading publisher in fields like Biology, with 119 books, and Social Science, with 103 books, demonstrating its strong presence in educational materials. On the other hand, Oxford was the top publisher in Maths, with 117 books, indicating its specialization in mathematical texts. Other publishers, such as Macmillan, Pearson, and Scholastic, also contributed significantly but varied in their focus across genres, with detailed counts showing each publisher's contribution to every genre category.
-        Visualization: To effectively communicate these findings, a bar chart can be used, with each bar representing a publisher and different colors assigned to distinguish between genres, allowing for an intuitive comparison of publishing volumes.

Objective 2: Book ID Search

- General Description: This objective involves creating a system to retrieve all comprehensive details of any specific book by using its unique Book ID, mimicking the functionality of a VLOOKUP in spreadsheet software, to ensure quick and accurate access to information.
- Specific Requirements: For any given Book ID, such as BK126, the system must extract and display all associated column values, including Title, Author, Genre, ISBN, Publisher, Publication Year, Edition, Pages, Shelf Code, Availability Status, Cost, Category, Book Condition, and Number of Reviews.
- Analysis Results: For example, when searching for BK126, the results showed that it is titled "NCERT Chemistry," authored by Madison Savage, belongs to the Chemistry genre, has an ISBN of 978-0-352-02233-2, was published by NCERT in 2016 in its 3rd edition, contains 374 pages, is stored in shelf E1, is currently available, costs 835, falls under the General category, is in average condition, and has 239 reviews. This level of detail ensures that users can access all necessary information swiftly and accurately.
- Visualization: The results can be presented in a single-row table format or as a detailed card layout, making it easy to read and understand all attributes at a glance.

Objective 3: Review-Wise Publisher Ranking

- General Description: This objective aims to rank all publishers based on the total number of reviews their books have accumulated, providing a measure of their popularity and influence within the library user community.
- Specific Requirements: The task involves summing up the "Number of Reviews" column for each publisher and then ranking them in descending order to identify which publishers are most reviewed and thus most impactful.
- Analysis Results: The analysis determined that NCERT achieved the highest total with 23,545 reviews, making it the most reviewed publisher, followed closely by Oxford with 21,569 reviews, Macmillan with 20,326 reviews, Pearson with 17,681 reviews, Scholastic with 16,950 reviews, and Cambridge with the lowest at 13,914 reviews. This ranking highlights NCERT and Oxford as the most popular among library users based on review counts.
- Visualization: A horizontal bar chart can effectively display this ranking, with each bar representing a publisher and its length corresponding to the total number of reviews, providing a clear visual comparison.

Objective 4: Shelf Visualization

- General Description: This objective focuses on analyzing and visualizing the distribution of books across their physical shelf locations, ranging from A1 to E4, to understand storage and usage patterns within the library.
- Specific Requirements: The analysis requires using interactive slicers to filter and explore the data, showing how many books are stored in each shelf, their availability status (Available or Issued), and other relevant details.
- Analysis Results: The results indicated that shelves like C3 and C4 contained the highest number of books, each with 40 books, reflecting high utilization in these areas. Other shelves, such as A1, had a mix of 16 available

and 14 issued books, while shelves like D3 and D4 showed varying levels of availability and issuance. This distribution helps identify which shelves are most and least used, informing potential reorganization or resource allocation strategies.

- Visualization: A stacked bar chart or a detailed heatmap can be used to visualize this data, with different colors or segments representing availability status for each shelf, enabling quick identification of patterns and trends.

Objective 5: Receipt Generation

- General Description: This objective involves developing a mechanism to generate formatted receipts for book transactions, ensuring that users and librarians have a clear record of borrowing or purchasing activities. - Specific Requirements: The system must allow input of customer details and book information to produce a receipt that includes all relevant data, such as customer name, book title, author, cost, and transaction dates. - Analysis Results: For instance, for BK126, a receipt would include customer information, book details like title "NCERT Chemistry," author, cost, shelf location, and availability status, formatted in a professional manner to serve as an official record of the transaction.

- Visualization: The output can be presented as a formatted text document or a digital template, designed to be clear, concise, and easy to print or save for record-keeping purposes.

5. Conclusion

The detailed analysis of the book dataset yielded significant and valuable insights into various aspects of library operations, including publishing trends, book availability, and user preferences. It was found that NCERT and Oxford stand out as the leading publishers both in terms of the volume of books published and the number of reviews received, clearly indicating their dominant position in providing educational materials that are highly valued by library users. The shelf analysis further revealed that certain locations, such as C3 and C4, are significantly more utilized than others, suggesting a need for strategic interventions to better distribute resources and optimize storage space across the library. Additionally, the ability to quickly search for specific books using their IDs and generate detailed, professional receipts for transactions greatly enhances operational efficiency, making library management more streamlined and user-friendly.

Overall, the dataset proves to be a robust and reliable foundation for effective library management, but it also highlights areas for improvement, such as addressing the condition of worn-out books, improving shelf organization to reduce overcrowding in high-use areas, and ensuring that resources are allocated in a way that maximizes accessibility and convenience for all patrons. These findings provide a solid basis for librarians and administrators to make informed decisions about future acquisitions, maintenance schedules, and user service enhancements, ultimately ensuring that the library continues to serve as a vital and efficient resource for its community.

6. Future Scope

The current analysis offers a strong foundation that can be further expanded and enhanced in several impactful ways to increase its utility and effectiveness for future library management. One key area for development is the integration of the dataset with a live, real-time library management system, which would allow for continuous updates on borrowing, returns, and availability, ensuring that the data remains current and actionable at all times. Another promising direction is the application of advanced predictive analytics and machine learning techniques to forecast book demand based on factors such as genre popularity, publisher reputation, and historical review trends, enabling proactive decision-making regarding stock replenishment and acquisitions.

Additionally, conducting a deeper analysis of user borrowing patterns could uncover valuable insights into user preferences and behaviors, which could then be used to develop personalized recommendation systems or targeted marketing strategies to enhance user engagement. The development of a user-friendly mobile application could also be a significant step forward, allowing patrons to search for books, check their availability, place holds, and generate receipts directly from their devices, thereby improving accessibility and convenience. Finally, a sustainability-focused assessment of book conditions, particularly identifying and addressing those labeled as "WORN," could inform strategies for timely replacements, repairs, or digitization efforts, reducing waste and extending the lifespan of library resources.

These enhancements would transform the dataset and its analysis into a more dynamic, interactive, and forwardthinking tool, paving the way for smarter, more efficient, and more user-centered library management practices in the future, ensuring that the library remains a cutting-edge and indispensable resource for all its users.