Data Mining Proposal:
Goodreads Book Data Analysis

Katelyn Stanton
Kennesaw State University
Marietta, Georgia, USA
kstanto7@students.kennesaw.edu

Abstract—**This project analyzes patterns in reader engagement and author productivity using data from Goodreads. The research aims to answer three discovery questions. Data mining techniques will be applied to uncover meaningful patterns in book and author attributes. The outcomes are expected to identify attribute combinations that drive reader engagement, highlight outlier books, and reveal clusters of authors based on productivity and quality metrics.**

I. Dataset

Dataset

Books Dataset GoodReads (May 2024):
https://www.kaggle.com/datasets/dk123891/books-datasegoodreadsmay-2024/data

Size

- Number of Columns: 15
- Number of Rows: 16,226
- File size: 22.3 MB

Data Description

This data set was compiled from Goodreads (May 2024) and contains 16,266 books with attributes that cover book data, reader engagement metrics, and quality indicators. The data set provides detailed information about books across various genres, publication years, and popularity levels

Key Features and Attributes

The primary attributes that will be used in this analysis include:

- Book Title
- Author's Name
- Genre
- Number of Pages
- Average Rating
- Number of Ratings
- Number of Reviews
- Publication Year

These attributes enable analysis of reader engagement patterns, book quality metrics, and author productivity across different generations and time periods.

II. Discovery Questions

Q1: What book attributes associate with higher reader engagement?

This question reveals which book characteristic(genre, length, author, popularity, etc.) consistently drive reader engagement through ratings and reviews. Understanding these patterns provides valuable insight for publishers and authors about which attributes attract reader attention and participation, which could help guide strategic decisions about book development and marketing.

Q2: Which books receive unusually high or low ratings compared to similar books in their genre and page length?

This question identifies outlier books that stand out from their typical ratings patterns within their category. These anomalies reveal components that influence reader satisfaction, providing insight into what makes certain books successful or unsuccessful.

Q3: What patterns in productivity and book quality distinguish successful authors from struggling authors?

This question explores what distinguishes successful authors from struggling authors by examining how book quality and productivity impact their careers. It analyzes whether successful authors prioritize writing fewer high-quality books or if they can maintain reader satisfaction while producing many titles.

Data Quality Issues

Potential quality issues may include missing values in certain attributes, duplicate entries for different book editions, and inconsistent genre tagging. Data cleaning and transformation will be conducted, including identifying and addressing missing values, removing duplicates, and standardizing categorical variables.

III. Planned Techniques

Q1: What book attributes associate with high reader engagement?

Technique: Association Rule
Relevance to Discovery Question:

Association rule mining will be used to identify which book attributes frequently occur together with high levels of reader engagement. Book characteristics such as genre, book length, and average rating will be transformed into categorical itemsets, and the FP-Growth algorithm will be applied to discover frequent patterns. This analysis will reveal combinations of attributes that consistently associate with high engagement metrics, including the number of ratings and reviews. The resulting rules will help identify which attribute combinations are most likely to drive active reader engagement.

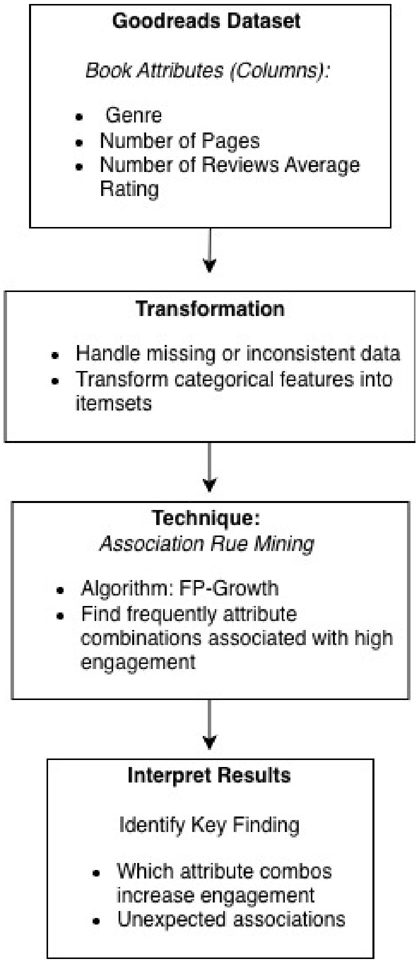
Goodreads Dataset
Book Attributes (Columns):
Genre
Number of Pages
Number of Reviews Average Rating
Transformation
Handle missing or inconsistent data
Transform categorical features into itemsets
Technique:
Association Rue Mining
Algorithm: FP-Growth
Find frequently attribute combinations associated with high engagement
Interpret Results
Identify Key Finding
Which attribute combos increase engagement
Unexpected associations

Fig. 1. Flowchart of the analysis for Q1

Q2: Which books receive unusually high or low ratings compared to similar books in their genre?

Technique: Anomaly Detection

Relevance to Discovery Question:

Anomaly detection techniques will be used to identify books whose ratings significantly differ from typical rating patterns within their specific genres. By comparing individual books based on attributes such as average rating and number of reviews, this analysis will highlight books that are unusually highly rated or poorly rated relative to their genre. This technique will uncover outliers that may represent reader reception and potential data inconsistencies.

Q3: What patterns in productivity and book quality distinguish successful authors from struggling authors?

Technique: Clustering

Relevance to Discovery Question:

Clustering techniques will be used to group authors based on shared patterns in productivity and book quality, using attributes such as number of books published, average book ratings, and total number of reviews. Algorithms such as K-means and hierarchical clustering may be applied to identify groups of authors with similar characteristics. These clusters will identify combinations of productivity and quality metrics that differentiate successful authors from struggling authors.

IV. Preliminary Timeline

M2: Initial Implementation

- Collect and clean the Goodreads dataset
- Transform and Clean Data
- Begin coding key analysis techniques for one of the discovery question
  - Q1: Association Rule (itemsets FP-Growth)

M3: Complete Implementation

- Run all planned data techniques on clean datasheet
- Continue coding key analysis for multiple discovery questions itemize
- Q1: Association Rule (itemsets FP-Growth)
- Q2: Anomaly Detection
- Q3: Clustering (K-Means / Hierarchical)

M4: Final Deliverable

- Complete all code for discovery questions
- Generate outputs and results for each questions
- Prepare presentation

Anticipated Challenges

- Interpreting clusters and association rules
- Learning and implementing new data mining techniques in Python
- Handling multiple anomalies
- Managing time effectively to complete M2-M4 milestones

References

[1] "*LaTeX*," Wikibooks, [Online]. Available: https://en.wikibooks.org/wiki/LaTeX. [Accessed: Feb. 5, 2026].

[2] M. J. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 1st ed. Cambridge, UK: Cambridge University Press, 2014.

[3] R. Raschka, "*MLxtend: Machine Learning Extensions*," [Online]. Available: https://rasbt.github.io/mlxtend/. [Accessed: Feb. 5, 2026].

[4] Scikit-learn developers, "*Clustering*," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html. [Accessed: Feb. 5, 2026].