# Relational Book and Audiobook Data for Comparison of Purchasing Factors for Readers

**Proposed By:**Lauren Sapp

**GitHub Repository:**[*https://github.com/laurenmsapp1/comparison\_books\_audiobooks*](https://github.com/laurenmsapp1/comparison_books_audiobooks)

Index

[Relational Book and Audiobook Data for Comparisons of Factors 1](#_Toc212662350)

[Project Overview 2](#_Toc212662351)

[Data Sources 2](#_Toc212662352)

[Dataset 1: 2](#_Toc212662353)

[Dataset 2: 2](#_Toc212662354)

[Dataset 3: 2](#_Toc212662355)

[Research Objectives 3](#_Toc212662356)

[Data Preparation Approach 3](#_Toc212662357)

[Current Status 3](#_Toc212662358)

[Deliverables (Remaining Work) 4](#_Toc212662359)

[Project Timeline 4](#_Toc212662360)

[Additional Considerations 4](#_Toc212662361)

## Project Overview

This project explores the relationship between physical and audio books across 3 platforms: Audible, GoodReads, and Google Books. By analyzing the data related to ratings, reviews, and cost, the project can provide insight to popularity and preferred method of entertainment. We will discover relatable information that would be valuable for a reader or publisher.

## Data Sources

### Dataset 1:

* **Dataset Name: Audible Complete Catalog**
* **File Name(s):** Audibly\_Catlog.csv, Audible\_Catlog\_Advanced\_Features.csv
* **Source:** Kaggle
* **Relevant fields**: Book Name, Author, Rating (Out of 5), Number of Reviews, Price
* **Link:** <https://www.kaggle.com/datasets/amritvirsinghx/audible-complete-catalog>

### Dataset 2:

* **Dataset Name: Google Books Dataset**
* **File Name(s):** google\_books\_1299.csv, google\_books\_dataset.csv
* **Source**: Kaggle
* **Relevant fields**: Title, Authors, Language, Rating, Voters, Price, Page Count, Publisher
* **Link**: <https://www.kaggle.com/datasets/bilalyussef/google-books-dataset>

### Dataset 3:

* **Dataset Name: Goodreads-books**
* **File Name(s):** Goodreads\_books.csv
* **Source**: Kaggle
* **Relevant fields**: Title, Author, Average Rating (Out of 5), Number of Pages, Ratings Count, Language
* **Link**: <https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks>

## Research Objectives

* **Primary Question(s):**
  + Cross Examination of Audiobook vs. Book Ratings- Is the book more enjoyable either way?
  + What is the relationship between the cost of the book and the rating?
* **Secondary/Exploratory Questions:**
  + What authors created the highest ratings?
  + Are readers recording that they have read a book on Goodreads?
  + What price point to readers feel comfortable spending on an audiobook vs. a physical book?
  + Does a date of publishing effect whether the book receives more ratings?

These values are interesting because they could be valuable to a publishing house deciding how much to set book cost for max sales. It could also contribute to decisions based around whether to invest in recording the book on audio. Goodreads provides giveaways for books and the amount of users recording rating could help decide the next book to distribute.

## Data Preparation Approach

* Align Dataset on books displayed on all 3 platforms. Create a source column on each data set prior to merging.
* Replace missing values with dropping rows unrelated to projection points. Additionally, rotate in information related to publisher from tables.
* Cap extreme values for cost
* Create composite indicators:
  + Rating Ratio: physical rating/audio highlighting comparable ratios
  + Popularity Index: Overall table view comparatively between all 3 tables.
* Store cleaned data in relational SQLite database with tables:
  + Book
    - Book\_ID (PK), Title, Author, Language, Page\_Count, Publisher, Source
  + Ratings
    - Rating\_ID (PK), Book\_ID (FK), Source, Rating, Number\_Reviews
  + Price
    - Price\_ID (PK), Book\_ID (FK), Platform, price

## Current Status

* Data acquired and inspected.
* Early cleaning or standardization steps.
* Prototype functions or exploratory plots.
* Evidence that the datasets can be successfully merged.

## Deliverables (Remaining Work)

* **Required Tasks (must be completed):***(Example: finish cleaning literacy dataset; implement 3 Python functions; build SQLite schema and load tables; produce 3 required visualizations; write README and data dictionary.)*
* **Stretch Goals (optional, if time allows):***(Example: add ACS demographics to analysis; build interactive dashboard; run regression model.)*

## Project Timeline

* Phase 1: Acquire and clean data, set up database.
* Phase 2: Exploratory analysis, build functions, first draft visuals.
* Phase 3: Deeper analysis, refine visuals, draft report.
* Phase 4: Finalize deliverables, polish repo, record presentation.

## Additional Considerations

Note any assumptions, limitations, or risks you anticipate. Examples:

* Geographic regions or variables you will exclude (and why).
* Additional datasets you may add if time allows.
* Possible challenges (e.g., very large files, inconsistent variable names, data sparsity).