# 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 into 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 number 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 publishers 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 from multiple platforms involving audio and physical books. (Google Books, Audible, and Goodreads)
* GitHub initialized/organized and ERD prepared for database construction.
* Prototype functions or exploration plots.

## Deliverables (Remaining Work)

* **Required Tasks (must be completed):**
* Continue to clean data and address adding columns to state platform source.
* Ensure naming is aligned for combination.
* Finalize SQLite Schema and load all four.
* Write and test three functions:
  + Engagement\_score() to combine popularity and usage.
  + Rating\_difference() to calculate the rating difference in physical and audio books.
  + Multiplatform\_engagement() to find total engagement throughout platforms.
* Generate three required visualizations:
  + Bar Chart to compare cost of books to the amount of time purchased (physical and audio) to determine a comfortable cost for readers.
  + Scatterplot to determine overall rating of titles per platform
  + Additional visual that is not determined at this time.
* **Stretch Goals (optional, if time allows):**
  + Group by Genre and determine if rating consistently appears across genres.
  + Work on consistency scores for authors who have multiple books on the table.
  + Review text of ratings for sentiment value.

## Project Timeline

* Phase 1: Acquire and clean data, set up database.
  + Finalize data clean up and prepare SQLite schema consisting of 4 tables.
  + Implement databases and functions.
  + Validate the joins and confirm book names aligned.
* Phase 2: Exploratory analysis, build functions, first draft visuals.
  + SQL breakdown to determine questions are usable for visuals
  + Produce draft versions of visualizations.
* Phase 3: Deeper analysis, refine visuals, draft report.
  + Finalize the required visualizations and ensure accurate depictions of question.
  + Draft README and polish structure of repository
* Phase 4: Finalize deliverables, polish repo, record presentation.
  + Finalize and organize all aspects of the repository (README file, requirements.txt, ERD Diagram, schema file)
  + Finalize all code with narrative and markdown clean up.
  + Submit Completed GitHub repository.

## Additional Considerations

Some of the data collected is data scraped from the platform. It is not the most accurate up to date but did consist of multiple books that matched up. The books that do not have an audiobook or physical book matching description will be data extracted but could be used in a dataset about the certain platform or way of acquiring book. If time allows, I will perform some additional testing on just the data for Audible for ratings. The data also does not show if the book was completed prior to rating.