

Louisville Libraries & GoodReads Books Outline

By Aliyah Gant, CodeYou Data Analytics Student

Objective/README file

The goal of this project is to compare physical book item inventory of the Louisville Metro (KY) Free Public Library system with GoodReads.com's physical book popularity ratings.

GoodReads is a website where readers can share critiques and compile personal lists of books that interest them (it does not sell or publish books).

Are there any trends or correlations between the GoodReads popularity of books/media and the Louisville Free Public Library's inventory?

Some people think public libraries are obsolete because of the rise of ebooks, but the Louisville Metro library system is bustling with patrons still, even for physical books. GoodReads is a website with 90 million international users, where readers can share critiques and compile personal lists of books that interest them (it does not sell or publish books). The hope is this project can be used to figure out whether Louisville is well-equipped for the global popular demand for certain books and media.

An assumption is that international GoodReads book/media demand may be similar to Louisville's book/media demand. Louisville Metro does not provide such details to the public, so there is not much to make an educated guess. This exercise will display final visuals that shed a light on any interesting trends or lack thereof.

How viewers may indulge:

- **Clone this repo.**
- **Setup the virtual environment:**
 - For Windows, run in your terminal:
 - `tutorial-env\Scripts\activate`
 - For Unix or MacOS, run in your terminal:
 - `source tutorial-env/bin/activate`
 - After you've finished looking at the project, type "deactivate" into your terminal.
- **You can check out `bookspandas.ipynb` for data aggregation & summaries.**
- **Open the Tableau document for the visual dashboard.**
- **Feel free to leave any comments in github!**

Methodology

The GoodReads data includes publications from 1/1/1980-12/30/2023 and their current ratings, reviews, "want to read" count, and "likes" count. The Louisville Metro library inventory data was compiled on 3/1/2024 and it gives basic detail of all items owned by the library system, even if they're currently being borrowed by patrons. All electronics and ebooks were filtered out from

both datasets. All 2024 publications in the library database were removed to allow more closely related time ranges.

Tables from both data sources were cleaned & merged into 3 relational tables - one describing the Books, another describing the GoodReads Popularity, and a third describing the Library Inventory.

The raw data used in this program was acquired from Louisville Metro Open Data and Kaggle.com via the below links.

<https://data.louisvilleky.gov/datasets/372216992aea4b2cb5b02837d7a48eaf/about>

<https://www.kaggle.com/datasets/cristaliss/ultimate-book-collection-top-100-books-up-to-2023/data>

Questions

1. How does the number of library copies compare to GoodReads popularity?
2. How many book copies does the Louisville Library system typically have of the more popular physical books on GoodReads?
3. Does the number of editions of a particular book available thru the library correlate significantly to GoodReads popularity?

Programming Methodology

1. Reading CSV files in via Pandas(Python) in Jupyter Notebook
2. Cleaning & Combining data via Pandas Merge, and using some DataFrame summaries.
3. Visualizations via Tableau.
4. Virtual environment to ensure compatibility.
5. Annotations via Jupyter Notebooks & README.

Data Cleaning Goals & Considerations:

- Created a relational database including a Books table, GoodReads Popularity table, & Library Inventory table.
- Added a total library copies column, removing the branch name column.
- Added a total versions/editions column to avoid ambiguity around different publications of the same book, removed duplicates of the same title & author as such.
- Removed other duplicates from both datasets.
- Removed ebooks & electronics from both datasets, to leave only physical books & audiobooks.
- Removed post-2023 publications from the library dataset.
- Removed irrelevant item types such as the library "Laptop".
- Cleaned up publication dates and author names to match across tables.
- Replaced blanks or invalid values when useful.
- *GoodReads Ratings >= Reviews by count, because users can rate without reviewing. But reviewers must rate.
- *Raw Library data - Each line is one copy of an item.
- *Raw GoodReads data - Each line is one recording of data (some duplicates).

Raw Louisville Metro Library Data Makeup

- Each line is one copy of an item
- Authors last name comma first name, some blanks
- Some blank ISBN's
- Some publication years look invalid (0, 120, 150)
- Some blank item collections/genres
- Some items with 0 price, "Laptops" with high price
- No index
- Louisville Open Data/Louisville Metro Government:
 - "Definitions: BibNum - The unique identifier of a bibliographic record within our materials database. Materials with the same bibliographic # will generally have the same cataloging metadata, differing only in the barcode number, assigned location and anything else specific to the individual copy."
 - <https://catalog.data.gov/dataset/louisville-metro-ky-library-collection-inventory-5e94f>

Raw GoodReads Data Makeup

- Some books are listed twice with everything the same except a slight difference in ratings or reviews.
- Some ISBN's are blank or start with "(ISBN" and are invalid
- Authors start with first name
- Some blank "formats"/media types
- Genres are in a "[,]" format with multiple genres per item
- Publication dates are in several different formats, some blanks
- 'Current Readers' has some decimals & some blanks
- 'Want To Read' has some decimals & some blanks
- Price has some 0's and some blanks
- Index starts at 0