Louisville Libraries & GoodReads Books Outline

By Aliyah Gant

Objective/README file

The goal of this project is to compare book/media item inventory of the Louisville Metro (KY) Free Public Library system with GoodReads.com popularity. 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? What about the market prices of the books/media?

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.

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 does item price compare to popularity?
3. How does item price compare to the number of library copies?

Programming Methodology

1. Reading CSV files in via Pandas(Python) in Jupyter Notebook
2. Cleaning & Combining data via Pandas Merge or SQL JOIN, and probably using some DataFrame summaries.
3. Visualizations via Matplotlib, Seaborn, Pandas, or Tableau.
4. Virtual environment and custom data dictionary.
5. Annotations via Jupyter Notebooks & README.

Data Cleaning Goals & Considerations:

- Most ideal to create a Books table and a Popularity table.
- Adding a total library copies column, removing the branch name column.
- Removing duplicates from both datasets.
- Removing post-2023 publications from the library dataset.
- Removing irrelevant item types such as the library "Laptop".
- Cleaning up item type labels to match across tables.
- Cleaning up publication dates to match across tables.
- Replacing blanks or invalid values in useful ways (more to come).

- Different publications of the same book can have different ISBN's.
- GoodReads Ratings >= Reviews by count, because users can rate without reviewing. But reviewers must rate.
- Library data - Each line is one copy of an item.
- 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

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