Joseph Petersen - October 2018 Data Science Career Track Cohort

Capstone 2 - First Milestone Report

Library Book Recommendation Engine

Overview:

John Albert's Library in Fictional Town, USA has recently added to their website the ability for library card holders to rate the books they check out, and in accordance with their mission to spread the joy of reading to their local community, they're interested in also providing recommendations for other titles based on previous ratings data. This project aims to create that recommendation system based off the Book-Crossing dataset.

Dataset and Methods:

Dataset: http://www2.informatik.uni-freiburg.de/~cziegler/BX/

The Book-Crossing dataset is comprised of three tables: BX-Users contains the user's ID as well as the location and age if available (null if not), BX-Books contains the ISBN, book title, book author, year of publication, and publisher, and BX-Ratings contains the book ratings information on a scale of 1-10 with some implicit values marked as 0. The dataset contains 278,858 users, 1,149,780 ratings, and 271,379 books.

The dataset will need to be cleaned up as there are some missing, null values, and questionable outliers. The user IDs and ratings will be formed into a matrix for user based collaborative filtering. This allows recommendations for users based off other users with similar reading habits and likes.

Data Wrangling Steps:

The first steps were to examine each separate table to look for errors and issues. The BX-Books table required the following wranglig steps:

1. Manual adjustments were made to the base file due to a number of extra separator characters (';') present in book titles. This was done by reading the errors flagged when setting the 'error_bad_lines' parameter in the Pandas read_csv function to 'False'. Additionally, '&' were coded in the table as '&' which were changed.
2. The book image URL's were removed as they are not necessary for the recommendation engine. The columns were also renamed for ease of use.
3. One null value in the 'author' column and two in the 'publisher' column were easily added with online information.

4. The ‘year’ column contained both integers and strings as well as two erroneous publisher names. The two names were easily examined and fixed with the proper entries as they just needed a shift over. The column was then converted to numeric. Some entries didn’t make logical sense (0 and years later than 2004, the year the dataframe was compiled), and they were converted to NaN.

The BX-Book-Ratings table required to following wrangling steps:

1. The columns were renamed for ease of use.
2. A large number of books (70,386) rated had ISBN numbers that were not present in the BX-Books table. These were removed as the title is necessary for recommendation.
3. The ‘ratings’ column contained 0’s which indicate an implicit rating meaning a type of rating that can’t be quantified on the scale of 1-10. These were separated from those with explicit ratings as the latter are what is required for the recommendation engine.

The BX-Users table required the following wrangling steps:

1. The columns were renamed for ease of use.
2. The ‘age’ column had entries that didn’t make sense. It was decided that ages less than 5 or greater than 100 were likely erroneous entries and replaced with NaN.
3. The ‘location’ column was split into city, state, and country.

Data Insights:

One important takeaway from looking into the books dataset was that some books appear under different ISBN. This is because of different editions. The next steps will need to include creating a second level of identification that groups together books with the same title but different ISBN.

Selected Poems	27
Little Women	24
Wuthering Heights	21
The Secret Garden	20
Dracula	20
Adventures of Huckleberry Finn	20
Jane Eyre	19
The Night Before Christmas	18
Pride and Prejudice	18
Great Expectations	17
Masquerade	16
Frankenstein	16
Black Beauty	16
The Gift	15
Emma	15
Beloved	15
Nemesis	14
Robinson Crusoe	13
The Wedding	13
The Secret	13

Table 1: The above shows the top 20 book titles by count. It indicates that many books have separate editions and therefore separate ISBN.

The books dataset also mostly contained titles published within the past 100 years. There is a distinct left skew. If publication year ends up playing a part in the recommendation system, this will need to be taken into account.

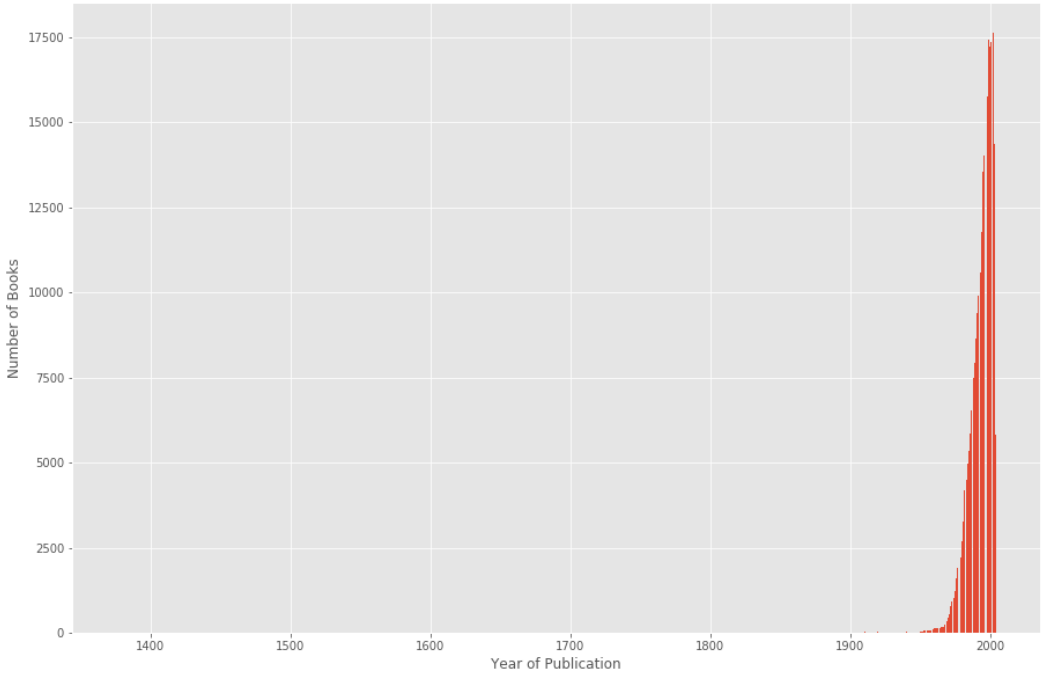

Figure 1: The above shows the heavy left skew of the distribution of publication years.

There's another skewed distribution when looking at the number of ratings per user. The vast majority of users only rate a few titles while a few distinct outliers rate more titles than most would even dream of. One user has rated over 14,000 titles. This indicates that the user rating matrix will be sparse (contains a lot of entries with no rating information). This was corroborated by checking the size of the user rating matrix, 121,052,287,740, versus the number of ratings, 1,149,780. Such a disparity indicates sparsity.

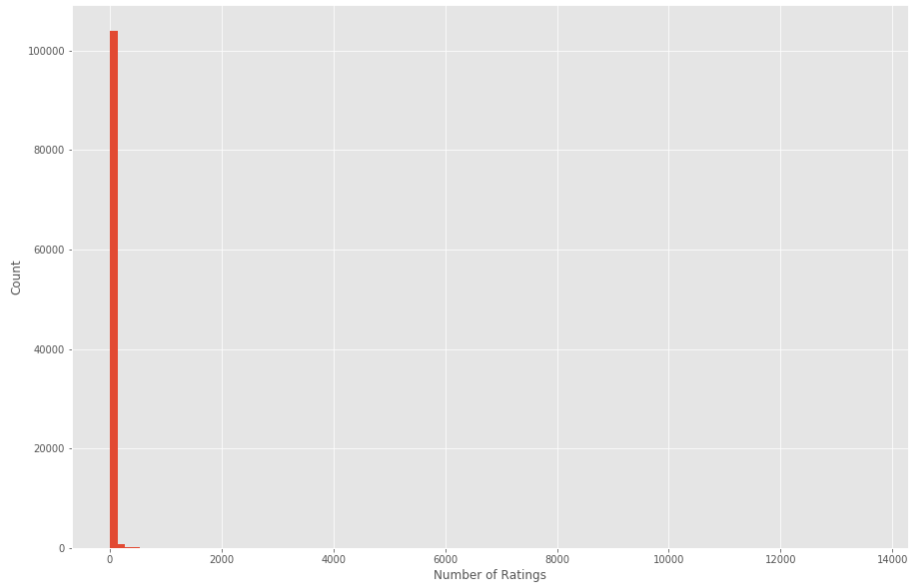

Figure 2: The above shows the incredibly heavy right skew in the distribution of number of ratings.

When looking at the counts of the ratings of 1-10, it showed that raters were generally favorable to the books they've read with the majority sitting over 5.

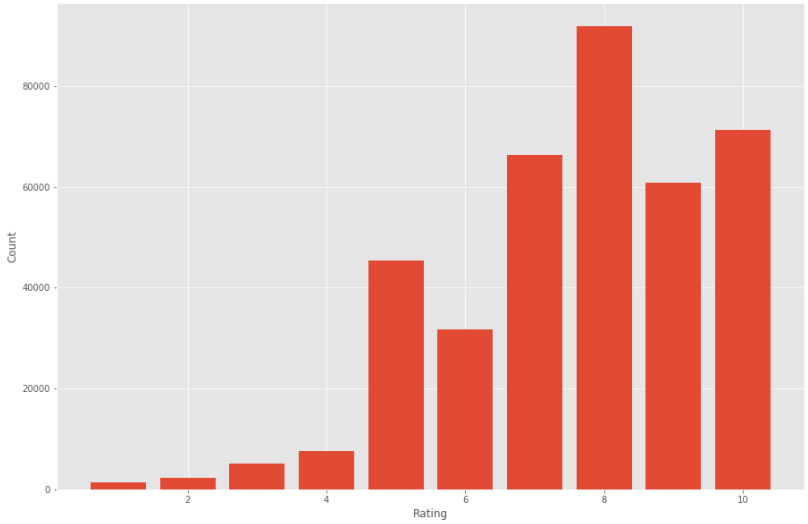

Figure 3: The above shows the counts of the ratings 1-10. It shows a general favorability by raters towards the books they've read.

The ages of raters slightly skewed right with the majority sitting around ages 20-40. This, unfortunately, does not take into account the large number of NaN values. Should age become a factor in the recommendation engine then a solution to those NaN values will need to be found. Likely, they would be imputed to the mean.

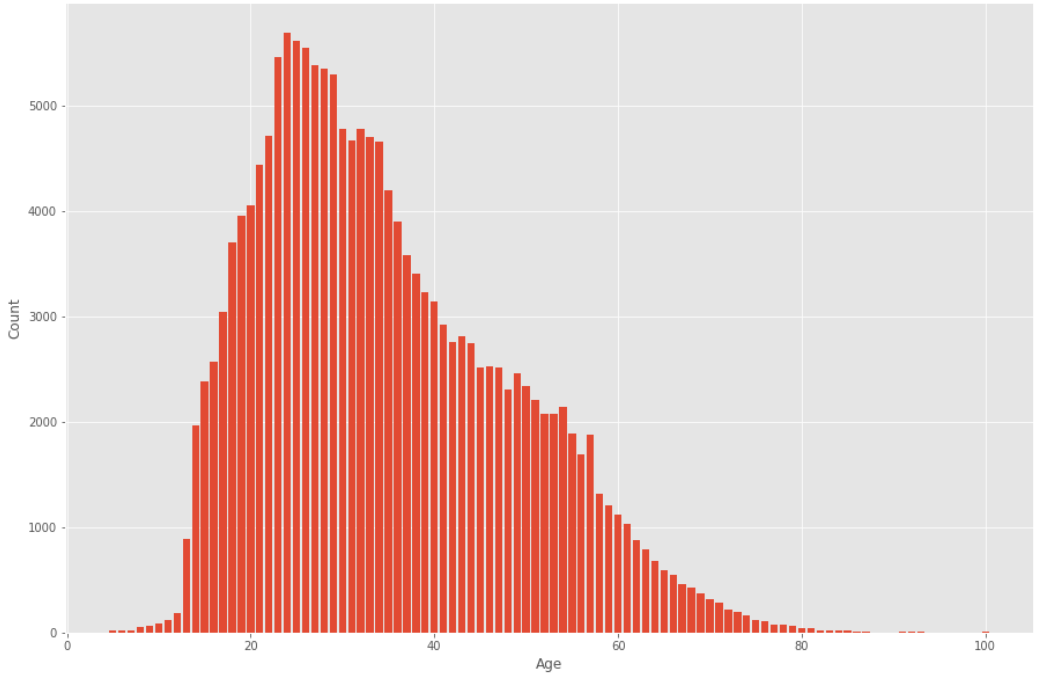

Figure 4: The above shows the distribution of the ages of users. The majority of the users lie between 20 and 40 years old.

The tables were joined on the BX-Book-Ratings table after the ISBN in it that did not correspond with ISBN in the BX-Books table were removed. It was a left join with BX-Book-Ratings as the caller due to those ratings being the target.

Next Steps:

Collaborative filtering is likely the best option to create the recommendation engine due to the explicit ratings in the dataset. The next steps will include testing on whether item-based or user-based collaborative filtering is the most effective.