Data Wrangling Project Overview

I obtained data from the following sources.

Flat Data Source

- Goodreads-books: https://www.kaggle.com/jealousleopard/goodreadsbooks

Website Data Source

- List of best-selling books: https://en.wikipedia.org/wiki/List_of_best-selling_books

API Data Source

- Open Library: https://openlibrary.org/dev/docs/api/books

The flat data source and the API have many relationships between them. I used the isbn13 from the flat data source pull the data from the API data source. The website data source shares the relationship of book title with the other two.

The CSV file is a comprehensive list of all books listed in Goodreads and was obtained from Kaggle. It has 11,127 rows of the following 12 variables:

- bookID – Identification number for each book
- title – Title of the book
- authors – Name of authors
- average_rating – Average rating the book received on Goodreads
- isbn – International Standard Book Number: 10 digit number unique to the book
- isbn13 – International Standard Book Number: 13 digit number unique to the book
- language_code – Primary language of the book.
- num_pages – Number of pages of the book
- ratings_count – Total number of ratings the book received
- text_reviews_count – Total number of reviews written for the book
- publication_data – Date the book was published
- publisher – Name of the publisher

The website data was obtained from Wikipedia and is a list of best-selling books, consisting of the following variables:

- Book – Name of the book
- Author(s) – Name of the authors

- Original language – Language in which the books was originally written
- First published – Year in which the book was first published
- Approximate sales – How much the book made in millions
- Genre – The genre of the book

The Open Library API contains over 20 million book editions with information about books. I used the isbn13 number from the flat data source to obtain additional information about the books. Some of the data extracted included title, genres, languages, publish_country, etc.

I performed data transformation and cleaning techniques on each of the datasets. Once each dataset was prepared, I merged and stored them in a SQLite database and created visualizations of my findings.